economic analysis
of law in india

economic analysis of law in india

THEORY AND APPLICATION

EDITED BY
P.G. BABU, THOMAS EGER, A.V. RAJA,
HANS-BERND SCHÄFER, and T.S. SOMASHEKAR

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

YMCA Library Building, Jai Singh Road, New Delhi 110001

Oxford University Press is a department of the University of Oxford. It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide in

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in India
by Oxford University Press, New Delhi

First published 2010

ISBN-13: 978 019 806056 7
ISBN-10: 019 806056 4

Typeset in Minion Pro 10.5/13 by Jojy Philip
Printed in India at De Unique, New Delhi 110 018
Published by Oxford University Press
YMCA Library Building, Jai Singh Road, New Delhi 110 001

CONTENTS

TABLES AND FIGURES

TABLES

FIGURES

ACKNOWLEDGEMENTS

The editors are grateful to the many people who made this project possible. In this respect we refer both to the conference in February 2008 at the Central University in Hyderabad as well as to the publication of the book. First of all, we would like to thank the European Commission, the University of Hamburg, the City of Hamburg, the Indira Gandhi Institute of Development Research in Mumbai, the University of Ghent, the National Law School of India in Bangalore, and the Central University of Hyderabad for financial support within the Asia Link project, 'Human Resource Development in Law and Economics for India and Europe'. In addition, we wish to thank everyone who supported us in organizing the conference and in the publication of this book. In this regard, we owe special thanks to Sönke Häseler, who reviewed some of the texts. Finally, we are truly grateful for the excellent and professional cooperation with the editorial team at Oxford University Press, our publisher, for all their assistance.

ACKNOWLEDGEMENTS

The editors are grateful to the many people who made this project possible. In this respect we refer to both the conference in February 2008 at the Central University of Hyderabad as well as to the publication of the book. First of all, we would like to thank the European Commission, the University of Hamburg, the City of Hamburg, the Indira Gandhi Institute of Development Research in Mumbai, the University of Ghent, the National Law School of India in Bangalore, and the Central University of Hyderabad for financial support within the Asia Link project, 'Human Resource Development in Law and Economics for India and Europe'. In addition, we wish to thank everyone who supported us in organizing the conference and in the publication of this book. In this regard, we owe special thanks to Sönke Häseler, who reviewed some of the texts. Finally, we are truly grateful for the excellent and professional cooperation with the editorial team at Oxford University Press, our publisher, for all their assistance.

INTRODUCTION

MOTIVATION

For much of her modern history since independence, India had a mixed economy in which the role of markets was limited. Large parts of the service sector and the industrial sector were state-planned and therefore required much discretionary power for governments and the bureaucracy. With increase in the number of markets, the importance of property law, contract law, business law, and corporation law has increased. The same applies to those fields of the law, like regulatory law, consumer protection law, and tort law, which correct failures of the market system. Fortunately, the rise of the market economy in India coincided with the rise of law and economics, a discipline which analyses the consequences of law for society, the conditions for legal reform, the scope for competition between jurisdictions, and the interrelationship between law and development by using the toolkits of modern economics. While this discipline does not reject the study of legal forms and their systematic extensions, which constitutes much of the traditional study of the law, it also insists that many areas of the law cannot be properly understood without using modern methods to evaluate their consequences for society. 'Law and Economics' is therefore an important discipline for lawmakers as much as for lawyers, and should be taught to law students as well as to students of economics and business administration.

We have little doubt that this discipline will flourish in India for various reasons, which arise from the demand as well as from the supply side. First, India is one of the few large developing countries in which the law in action, adjudicated by judges and enforced by the state, plays a crucial and increasingly important role, notwithstanding

some well-known shortcomings. Businesspeople and consumers increasingly rely increasingly on anonymous partners in anonymous markets such as the stock market and the insurance market. Moreover, they increasingly trust the law and its sanctions against opportunistic behaviour of partners, predators, and invaders, rather than doing business exclusively with friends, family, and other close and trust worthy persons. When one of the authors visited India 30 years ago, packed or canned food was not available for purchase; milk was bought directly from the owner of the cow. All this has completely changed and packaged food is now available nationwide. The rapidly growing stock market is an indicator of strong confidence in the corporate governance structures on the part of both, Indian and foreign investors. The law has helped tremendously in making this happen. This trend creates a huge demand for improving the law and for making it functional to support anonymous transactions. The discipline of law and economics can accompany this process in many fields of Indian law.

Second, India has a legal system in which the Supreme Court plays a fundamental role in shaping the country's future, comparable to the US Supreme Court or the European Court of Justice. Such courts have to rely on dependable information about the social consequences of high court decisions. Any legal culture in which Supreme Courts develop the law by judge-made rules has a need for legal scholarship and academic discussion that focus on the consequences of alternative legal solutions to a problem.

Third, academic talent abounds in India, especially in the field of economics. The Nobel prize for Amartya Sen is only the tip of the iceberg. A sizeable number of Indian economists now work in the fields of law and economics, law and development, and institutional economics. They have published books and articles in international and widely-read journals. They have presented concise theoretical and empirical studies on law and development, corporate governance, contract law and its substitutes, tort law, consumer protection law, securities regulation, environmental law and more. A book like this one is therefore not intended as a starter but as a contribution to an ongoing debate.

Though several scholars such as T.C. Anant, Kaushik Basu, and Dilip Mookherjee have made fundamental contributions to the theory

and practice of Law and Economics in India, the time has come to give further momentum to the debate. Some initiatives have already been taken, the most prominent of which are the Center for the Study of Law and Governance at Jawaharlal Nehru University Delhi, the Delhi School of Economics, and the Indian Institute of Technology at Kanpur. Moreover, the ongoing collaboration between the Indira Gandhi Institute of Development Research (IGIDR), Mumbai, and the University of Hamburg over the past eight years has culminated in IGIDR joining the Erasmus Mundus Masters Programme in Law and Economics as the first Asian institution. We are witnessing a heightened sense of awareness and usefulness of Law and Economics among Indian law students and legal scholars, thanks to the advent of a series of National Law Schools established in Bangalore, Hyderabad, Kolkata, Jodhpur, Chandigarh, Lucknow, Gandhinagar, and Delhi, and private law schools such as Jindal Global Law School, whose undergraduate and masters curriculum explicitly focus on the interrelationships between law and other social sciences. Interestingly, the National Judicial Academy at Bhopal, an institution established to train judges, guided by the Supreme Court of India, has initiated a course for judges in economic methodology. Last but not least, the *Indian Journal of Law and Economics* was recently launched in 2009, and has two of this book's editors on its advisory and consulting editors' boards.

THE ASIA LINK PROJECT

This book is the result of a larger cooperation between the Indira Gandhi Institute of Development Research (Mumbai), the Central University of Hyderabad, the National Law School of India University (Bangalore), and the law faculties of the Universities of Ghent (Belgium) and Hamburg (Germany). This cooperation enjoyed three years of support from the European Commission within the 'Asia Link' framework and essentially comprised three activities.

First, law and economics scholars from Belgium and Germany conducted four student workshops in Mumbai and Bangalore, in which they taught courses in different fields of Law and Economics to more than 160 Indian students. Second, as part of a 'sandwich programme', five excellent Indian PhD students who were selected

by the Indian partner institutions spent around 20 months at the Universities of Hamburg or Ghent, under the supervision of German or Belgian professors of Law and Economics, respectively. Third, a senior scholar workshop for Indian and European scholars on 'Law and Economic Development' took place at the Central University of Hyderabad, organized by Angara Raja (local organizer) and Thomas Eger (Head Coordinator). This volume comprises most of the (revised) contributions to the Hyderabad workshop, complemented by two additional papers by P.G. Babu and Robert Cooter and Hans-Bernd Schäfer.

The senior scholars' workshop in Law and Economics at the Central University of Hyderabad, on 28–9 February 2008, brought together researchers from different fields of Law and Economics who addressed a number of broad issues of relevance to the Indian economy. This book is divided into several sections, each of which focuses on a particular subject. Thus, Part I deals with financial institutions; Part II with accident laws and environmental problems; Part III tackles trade agreements and intellectual property rights protection; Part IV, corruption and its effects; Part V, the role of the judiciary in fostering efficient enforcement of law; and Part VI addresses the impact of non-legal institutions on economic development.

With regard to Part I, the law and finance literature shows that effective creditor and investor protection are a prerequisite for vibrant financial markets. The Finance and Growth literature in turn establishes a strong causation between a robust financial sector and economic growth. In most developing countries, the problem of finance takes on two important dimensions, the first being the impact of weak laws and a weak legal system on the allocation of credit by financial institutions, and the second issue being the degree of credit availability to the rural sector. Without effective legal protection, potential lenders would reduce their credits and require higher risk premia—which negatively affects the accumulation and allocation of capital and the rate of economic growth. Regulation is the answer to most problems but is in turn accompanied by its own set of problems.

The contribution by Robert D. Cooter and Hans-Bernd Schäfer on 'The Double Trust Dilemma: Reflections on Law and Growth' reflects on the difficult relationship between innovators and investors to the

law. Compared to sustained per capita growth, other sources of wealth are insignificant for overcoming mass poverty. Economies enjoy sustained growth through innovation. Growth through innovation faces a double trust dilemma. The innovator must be trusted by a financier without disclosing his innovation and vice versa. Contract law, property law, and business law can overcome this dilemma. Some people argue that poor countries do not face this problem because they can simply follow the rich countries and imitate their ideas. The double-trust problem is therefore a problem of rich countries and not of poor countries. We believe that this is wrong. If it were true, the poor countries would be rich already. Strategies to overcome static reproduction may be different in poor countries, but they are needed as much in poor countries as in rich countries. In order to convince people in India to buy edible oil in cans and packed food, and to trust the producer so that markets may expand in time and space requires no less of a good idea than to find a new marketing strategy in a rich country.

In his paper on 'Unit Linked Insurance Policies: Do Regulatory Gaps Pose a Moral Hazard?', T.S. Somashekar focuses on regulatory lags in the Indian insurance sector due to rapid changes in product profiles. With the development of the life insurance sector, and the marketing of Unit Linked Insurance Products (ULIPs), which combine features of mutual funds and insurance, the regulator, the Insurance Regulatory and Development Authority (IRDA), has been found wanting. Weak regulations covering portfolio disclosure, expenses, and agent commissions, compared to similar regulations for mutual funds by the Securities and Exchange Board of India (SEBI) have led to moral hazards. Portfolio selection skills, on an average, are found to be poorer for ULIPs. Furthermore, their higher levels of commissions for agents have created an uneven playing field where agents have an incentive to sell ULIPs rather than mutual funds despite the latter delivering better returns and being better regulated. This could lead to sub-optimal choices on the part of investors. The paper points to the need for regulatory bodies to be fleeter footed and, probably, to the need of a super-regulator at a time when products overlap substantially.

The third paper, by Alexander S. Kritikos on 'Repayment Behaviour under Joint Liability Loan Contracts in a Competitive Environment'

analyses the efficacy of micro-lending contracts in changing environments. More specifically, Kritikos investigates the impact that competition between microfinance institutions may have on joint-liability approaches, particularly on the peer selection process, and the repayment behaviour of groups. He finds that competition does not only result in positive effects: there is preliminary evidence that the indirect costs of group-lending contracts, such as the cost of monitoring and enforcement, increase because the self-selection processes of peer groups do not evolve to the same extent as they would in a non-competitive environment.

The essays in Part II, 'Accident Law', pose a few questions: Every country in the world has adopted laws that regulate the environment and protect the citizens from accidents. Why do countries differ on these laws? Should different countries have different standards for accident laws? Of the two papers related to accident law, the first one is a theoretical paper that examines liability rules in case of accidents, whereas the second one discusses a specific Indian example.

The paper by Satish K. Jain on 'Negligence Rule: Some Strategic Aspects' discusses the negligence rule as an efficient liability rule. An efficient rule minimizes the sum of precaution costs and expected accident loss or, in short, the social cost. This is possible, he points out, if both the victim and the injurer take levels of care that minimize social costs and the more able persons take a higher level of care. He examines the incentive effects of three types of negligence rules, namely the due care levels fixed on a case-by-case basis, a uniform due care level, and cost-justified levels of precaution. All three are found to be subject to strategic manipulations by the least-cost bearers either individually, or as a group. They have an incentive not to increase the effectiveness of their care or to invest in misleading the courts about their care-taking abilities and their ability to foresee risk. His finding has important implications, as the scope for strategic behaviour tends to upset the social cost minimization objective of the negligence rule.

The second paper by P.G. Babu on 'Suit and Settlement under Asymmetric Information: The Case of Bhopal Gas Disaster' models and analyses the famous suit between the victims of the Bhopal gas tragedy and the Union Carbide Corporation. The analysis relies on insights from information economics and bargaining theory. It

corroborates theoretically predicted conflicts between individually rational actions/choices on the one hand, and the collectively rational choices on the other hand. Additionally, the paper highlights the conflicts among the objectives of the parties involved, which are akin to agency relations.

India is a founding member of the General Agreement on Tariffs and Trade (GATT) of 1947 and its successor, the World Trade Organization (WTO), which came into effect on 1 January 1995, after the conclusion of the Uruguay Round (UR) of multilateral trade negotiations. India's participation in an increasingly rule-based system in the governance of international trade is to ensure more stability and predictability, which ultimately should, according to the proponents of free trade, lead to more trade and prosperity for itself and the 134 other nations which now comprise the WTO. India also automatically avails itself of any most-favoured nation clauses and national treatment for its exports to all WTO Members. Perhaps no other initiative has provoked such heated debate in the country as India's joining the WTO. The issues addressed in Part III, 'Trade Agreements and Intellectual Property Rights' range from farmers' rights to food security, patents, and the future of trade. The two papers on this subject address two issues that are of particular importance: the nature of regional trade agreements in the light of the WTO agreements, and the new patent regime and its impact on innovation in India.

In his article on 'Regional Trade Agreements: Building Blocks or Stumbling Blocks' A. Jayagovind argues that while it may be necessary to leave room for regional aspirations (Article XXIV recognizes and provides for such aspirations), most of the agreements in place are hardly 'regional'. He states that the numerous overlapping regional agreements could threaten multilateralism. The basis for this lies in the ineffective and unclear WTO regulatory and institutional framework, which comprises the Committee of Regional Trade Agreements (CRTA) and the General Council (GC). The CRTA examines proposals for the formation of Regional Trade Agreements (RTAs) and forwards its reports to the GC. Despite the GC not accepting most of these reports, countries have gone ahead with their own agreements. The role of the Dispute Settlement Board in pronouncing rulings on RTAs is also problematic as it involves areas where the contracting

parties have not reached a consensus. Developing countries face the peril of unequal bilateral negotiations with economically stronger, developed countries. This problem is compounded further as all agreements between developed and developing countries are covered by the Preferential Trade Agreements (PTA) provision, which has less stringent obligations. Given the lax regulatory system this has frequently been abused by developed countries to impose non-trade or contentious issues such as labour rights, environmental standards, and stricter intellectual property right (IPR) standards on developing countries. The author feels that the much abused and ineffectively regulated PTAs or RTAs need to be re-examined, and emphasis placed directly on multilateralism.

Thomas Eger, Petra Ebermann, and Padmanabha Ramanujam discuss in their paper on 'Incremental Innovation and Patent Protection for Pharmaceutical Products in India: A Law and Economics Analysis of the *Novartis* Case' the recent court decision of the Madras High Court against Novartis. In order to comply with obligations under the WTO Agreement on Trade-Related Aspects of Intellectual Property Rights, India has enacted the Patent Act of 2005 which allows for product patents in the pharmaceutical sector. With Section 3 (d) of the Act, an exemption has been introduced which excludes improvements without an increased efficacy from patentability. The Madras High Court's rejection of Novartis' challenge of Section 3 (d) has re-opened the debate on whether disallowing patents on incremental innovation or follow-up innovation will discourage firms from doing fresh research. This paper discusses how the scope of patent protection can be determined in a way that a competitive environment for improvements or incremental innovation is generated without extensively reducing the pioneer's incentives.

One of the main reasons for the lack of faith in the legal system to protect the citizens from predatory practices is corruption. This problem is discussed by Ajit Mishra in his paper 'Enforcement and Corruption: Some Implications of "Soft" Information', which makes reference to recent developments in India. Corruption leads to severe dilutions and a distortion of the regulatory and legal enforcement process. The paper argues that in many low-income societies information can be easily manipulated and access to judicial process is

relatively costly. It shows that in the presence of soft information and high judicial costs, societies can get caught in a low compliance trap and optimal policies lead to excessive under- or over-enforcement.

The role of the judiciary is crucial in fostering good institutional practices by efficiently dealing with cases that violate laws. Part V illustrates that while corruption in the judiciary is an important impediment, there are at least two other dimensions to the problem of an effective judiciary: (i) the role of the judiciary in a democracy, and (ii) the problems of an overburdened judiciary.

The first problem is concerned with what is known as judicial activism. The Supreme Court of India has become a policy maker of a whole range of issues such as human rights, environmental affairs, the functioning of political parties, and municipal agencies. This kind of activism also varies in character: interpretational activism, legislative activism, or executive activism. Public Interest Litigation (PIL) is one such example where the Supreme Court of India (SCI) has widened its scope and jurisdiction by interpreting the constitution to expand the notion of fundamental rights and by relaxing the requirements of locus standi. There is convincing empirical evidence to show that public interest litigation in environmental cases is at least partially effective in India. The paper by Michael G. Faure and A.V. Raja on the 'Economic Analysis of Public Interest Litigation in Environmental Cases in India' systematically studies the key variables that make public interest litigation successful in India. One of the important conclusions is that when account is taken of capacity and corruption problems, given the presumed information advantage of regulatory authorities (which are unfortunately common in developing countries), a second best solution may be chosen which deviates from the traditional advice provided by law and economics scholarship. Apparently the (relative) success story of public interest litigation in India fits into this strand of literature. In those situations it may be better to look for second best solutions if other levels of government or instruments are less vulnerable to the problems mentioned.

The second aspect of the judiciary under focus is that of judicial delays and their causes. R.V. Ramana Murthy and Siddik Rabiyath deal in their contribution on 'Disposal Rates, Pendency, and Filing in Indian Courts: an Empirical Study of Two States' with the crucial problem of

judicial delays in India. With a backlog of almost 30 million cases in the high courts and lower courts, litigants' transaction costs increase significantly due to the associated problems. The inability of courts to resolve disputes in a timely manner has important consequences for economic growth and development. The authors examine the impact of changing court efficiency, which is defined as disposal rates, on the rate at which fresh cases are filed. An improvement in efficiency would increase confidence in the courts, and therefore tend to increase the rate at which people approach them for justice. Studying data from two states in India, the authors find a positive correlation between changes in efficiency and the rate at which fresh cases were filed. While one state witnessed declining efficiency, it also saw a fall in the rate at which fresh cases were filed. This finding is worrisome as it possibly indicates out-of-court settlements under duress. There can be no certainty of optimal outcomes. The authors suggest increasing the number of judges to improve disposal rates.

Finally, in Part VI, the last paper in the section addresses 'The Impact of Non-Legal Institutions on Economic Development'. The contribution by T. Ulen on 'The Limited Role of Law in Economic Development' is a survey of the issues and the literature on the role of legal and other institutions in economic growth. The author has countered the popular hypothesis that good laws play a significant role in economic growth. The author's arguments focus on the following three points. First, that there are non-legal institutions that can be a substitute for 'good' laws. Second, social and political stability play a crucial role for growth. Good laws may or may not provide stability and hence may not be crucial for growth. Third, visionary leadership is a precondition for economic growth of developing economies.

I

FINANCIAL INSTITUTIONS

1

THE DOUBLE-TRUST DILEMMA IN DEVELOPMENT

Robert D. Cooter and Hans-Bernd Schäfer

1.1 SEPARATION OF IDEAS AND CAPITAL

When someone discovers a better way to make something or something better to make, developing the new idea requires capital. New ideas and capital repel each other like magnets with the same charge. An example illustrates the problem: An economist who worked at a Boston investment bank received a letter that read: 'I know how your bank can make 10 million dollars. If you give me 1 million dollars, I will tell you.' The letter captures concisely the problem of financing innovation: The bank does not want to pay for information without first determining its worth, and the innovator is reluctant to disclose information to the bank without first getting paid. As depicted in Figure 1.1, the separation of ideas from capital is a problem of double-trust: The innovator must trust the investor not to disseminate the idea, and the investor must trust the innovator not to appropriate the capital.

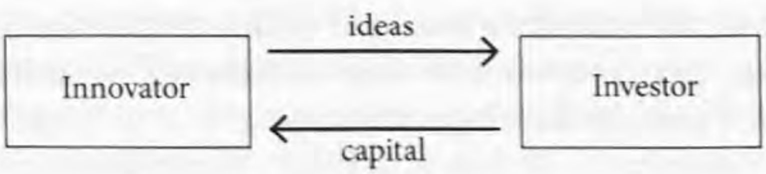


FIGURE 1.1: Double Trust Problem

Source: Author.

Note that the double-trust problem of growth is conceptually different from the problem of growth by capital accumulation. If the capital stock per worker increases in an economy it leads to a higher product per worker and therefore to per capita growth. But capital accumulation is possible in a world without any new ideas where the double-trust problem does not arise.

Capital accumulation can raise labour productivity to a limited extent and is subject to the law of diminishing returns. A farmer with a wooden plough produces less wheat than if he used a steel plough and a horse and he can even produce more if he uses a tractor and chemical fertilizer. This capital accumulation makes use of known technologies, not of new ideas, and its contribution to growth is very limited. The heart of the matter is to realize new ideas, which leads to the double-trust problem. Some people argue that the double-trust problem is faced only by rich countries; poor countries do not face this problem because they can simply follow rich countries and imitate their ideas. We believe that this is wrong: if that was the case, the poor countries would be rich already. The new ideas to break away from static reproduction are different in poor countries, but they are as much needed there as in rich countries. To convince people in India to buy edible oil in cans and packed food and to trust the producer and therefore to extend the market in time and space requires no less of a good idea than to find a new marketing idea in a rich country for the Nike corporation.

1.2 THE DOUBLE-TRUST DILEMMA

How is the double-trust dilemma solved? In the past, to secure peace between two hostile states, each one gave the other something of value as hostage. In antiquity, King Geiserich of the Vandals who ruled Cartago (now in Tunisia) offered his son as hostage to rulers of parts of North Italy King Theoderich of the Visigoths, who in turn gave Geiserich as a show of trust, his daughter as hostage.[1] Hostage exchange works best when each side values cooperation more than the hostage that it receives. In these circumstances, each side loses more than it gains by harming its hostage.

[1] The son and daughter were to marry if peace was preserved. Alas, Theoderich allegedly plotted against Geiserich, so Theoderich's daughter was mutilated and sent back to her father.

A hostage in war resembles a bond in business. To develop an innovation, the innovator explains his secrets to the investor and the investor supplies money to develop the innovation. A double bond can solve the double-trust dilemma of economic growth. If the investor keeps the secrets and the innovator uses his best efforts to develop the business, both expect to enjoy extra-ordinary profits. Alternatively, if the financier steals the secrets of the innovator, or the innovator appropriates the financier's investment, their new business will fail. The investment is the financier's bond that s/he risks losing if s/he fails to keep the innovator's secrets. The secrets are the innovator's bond that s/he risks losing if s/he fails to use the investment to develop his ideas. The double bond is effective as long as each side believes that collaborating to develop the innovation will produce more value than any alternative use of the secrets and the money. The two interlocking rings in 'Solomon's Knot' or 'King Solomon's Knot of Wisdom' stand for the bond between innovator and financier that needs to be unbreakable.

1.3 INFORMATION DIFFUSION AND SILICON VALLEY FINANCE

We will show that double bonding in business requires law. First, however, we must describe the life cycle of an innovation. Like the secret recipe for Coca Cola, private information belongs to a few people. An innovator has private information that other people lack. If the innovation is useful, the innovator gains a competitive advantage from its development. The competitive advantage yields exceptional profits, which attract competitors who try to learn what the innovator knows. Thus Coca Cola attracted competitors like Pepsi who learn to make similar drinks. As competitors come to understand what the innovator knows, the innovator's private information becomes public. Public information is available to everyone, like chemistry in a textbook. As an innovation becomes public, the innovator loses his/her competitive advantage and his/her profits return to the normal level.

This suggests three stages in an innovation's life cycle: First, someone has a new idea and obtains capital to develop it. The innovator may form a new firm or work inside an established firm. At the first stage, only a few people in the innovator's inner circle understand the innovation.

Here, the innovation's economic value has not been established. First-stage investors take the greatest risk of failure and earn the highest profits in the event of success. Next, the innovator develops the innovation sufficiently for it to succeed in the market. The innovator's organization enjoys exceptional profits and it expands faster than its competitors. Although profits are exceptional, second-stage investors take less risk and earn lower profits than first-stage investors. Thereafter, competitors observe the innovator's success and try to discover what the innovator knows. As competitors emulate the innovator, the innovator's profits fall and the innovation's growth slows. In the end, competitors assimilate the innovation, and the innovator's profits and risk return to an ordinary level. Decreasing privacy, risk, and profits characterize the three stages in an innovation's life cycle, as depicted in Figure 1.2.

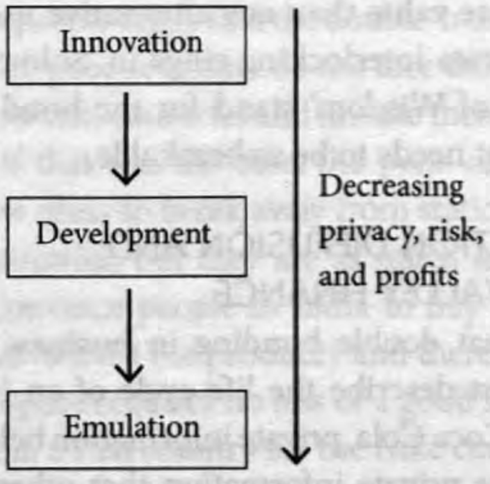


FIGURE 1.2: Life-cycle of an Innovation

Source: Author.

The three stages in an innovation's life cycle correspond to three phases of finance in Silicon Valley, as depicted in Figure 1.3. These are: relational, private, and public finance.

According to a popular quip, initial funding for startup firms comes from 'the 3 Fs': family, friends, and fools. Family and friends have confidence in the innovator, even though they cannot evaluate the innovation. Personal relationships motivate these investors, so we refer to the first stage as *relational finance*. A few fools may invest who think that they can evaluate an innovation without understanding it in such ventures. (Later we explain why government officials may not necessarily do a better job.)

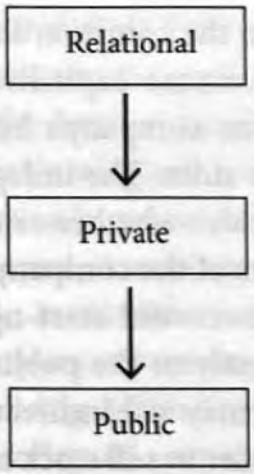


FIGURE 1.3: Finance in Silicon Valley

Source: Author.

Most innovators have too few personal relationships with wealthy people to finance an innovation's full development, so they must eventually turn to strangers. The second stage of funding comes from 'venture capitalists' who are not family, friends, or fools. Unlike relational finance, venture capital is a form of *private finance*. Finance is private because it comes from a small group of experts at evaluating innovations that are in an early stage of development. The problem of trust is severe between founders and venture capitalists. The creative leads who found a company often manage it badly. When the founders prove to be bad managers, the venture capitalists must take the company away from them and put its resources under the command of good managers. However, venture capitalists may use their power to seize the firm in order not to share profits with the founders, even when they are good managers. Besides 'venture capitalists', the initials v.c. can also stand for 'vulture capitalists'.

To solve the double-trust dilemma, the founders commit to performance goals in exchange for financing. If they fail to meet the goals, they lose their investment and their jobs, because the preferred shareholders (who are the venture capitalists) wipe out the common shareholders (who are the founders) and appoint new management. The financing contract may say that preferred shareholders can demand repayment of their investment within three years, so the founders must earn enough profits to repay the venture capitalists within that period

or risk losing the firm. Or, the common shareholders (founders) and preferred shareholders (venture capitalists) may appoint an equal number of directors to the company's board, plus an independent director accepted by both sides. The independent director holds the decisive vote and thus decides whether or not to replace the founders and appoint new managers of the company.

In the third stage, a successful start-up sells itself to the public. The start-up may sell directly to the public through an initial public offering of its stock, or it may sell indirectly when a publicly traded company acquires it. In order to sell stock to the public in India a firm must comply with disclosure rules of the securities laws. After the information is disclosed, brokers quickly disseminate it to the investors whom they advise. Many people understand the innovation sufficiently to decide whether or not to invest in its further development. Because investors are a large group of people, we describe the third stage as public finance.

Notice the tight connection between information as depicted in Figure 1.2 and finance as depicted in Figure 1.3. Only the innovator understands the innovation in the first stage and finance is relational. A small group of specialists understand the innovation in the second stage and finance is private. Many people understand that the idea in the third stage and finance is public.

Like biological mutations, most startups fail and a few succeed spectacularly. The first stage is the riskiest for investors; so first-stage investors get the stock at the lowest price. Those startups that survive the first stage have something substantial to show to venture capitalists in the second stage. The innovation is easier to understand, but it has not proved its profitability in the market, so second-stage investment is still very risky. Venture capitalists get the stock cheap, but not as cheap as first-stage investors. Finally, having proved its profitability in the market in the second stage, a successful firm ceases to be a startup when it reaches the third stage. Third-stage investors face lower risk and pay a higher price for stock. In each stage of finance for a successful company, the risk and profits decrease, and the stock's price increases. In the stage of public finance, the firm tends towards an equilibrium in which information and finance become public, so the double-trust dilemma dissolves.

1.4 ENTREPRENEURIAL INNOVATION AND FINANCE

In every country, growth occurs through innovation, but the form of innovation differs with the level of development. In Silicon Valley, innovations are primarily technological, such as new computer chips or programs that were previously unknown to the world. Technology mostly flows from developed countries to developing countries through international trade, investment, and educational exchanges. By the end of the last century, most international obstacles preventing poor countries from acquiring technology disintegrated as major wars ended, communism collapsed or transformed itself, and tariffs and transportation costs fell. The remaining obstacles to acquiring technology are mostly domestic, notably bad organization and impediments to markets.

Instead of innovations in technology, many innovations are 'entrepreneurial,' by which we mean that they improve organizations and markets. Philip Knight, co-founder of the Nike Corporation, began by selling running shoes out of the trunk of his car. From this beginning in 1972, the company reported 15 billion dollars in worldwide sales of sports equipment and clothing in 2006. Knight obviously discovered something new, but what was it? The company does not manufacture anything. Its main facility in Beaverton, Oregon, is a 'campus,' not a factory. Instead of manufacturing, it contracts with foreign companies to make all the goods that it sells. The business of Nike is research and marketing. Its research discovers new products, which it contracts with other firms to make, and then it markets them through extensive advertising. This new organizational form has spread dramatically in America as more and more companies 'outsource' manufacturing and focus on research and marketing. Other examples of recent innovations in organizations and markets in the United States include hostile takeovers of public companies, debit cards, extended warranties on complex consumer products, and team production in manufacturing (imported from Japan).

Developing countries lack top universities and large research organizations; this limits the scope of technological innovation. Entrepreneurial innovation, consequently, is more important to their economies than technological innovation. Entrepreneurs in developing countries, like Leo Stan Ekeh from Nigeria, who found a

way to produce computers in Africa, often transfer organizations and markets that originate in developed countries and adapt them to local conditions. Adapting organizations and markets to local conditions removes the domestic obstacles to absorbing foreign technology and increasing productivity of workers. Adaptation in developing countries faces the same obstacles as invention in developed countries: this idea requires capital and is a risky venture. The success of such a venture fuels competition, which, in turn, diffuses the innovation, thereby reducing its profit margin. Here, too, the double-trust dilemma holds time: the innovator and the investor must trust that they will come through for each other.

Biologists sometimes say 'ontogeny recapitulates phylogeny', which means that the development of a single organism from birth to maturity resembles the evolution of the entire species. Similarly, the three stages of finance for a start-up in Silicon Valley resemble three stages of historical evolution of capital markets for countries. First, in countries without capital markets, businessmen mostly borrow from family and friends. Finance remains relational, which keeps businesses small and local. Some peoples, notably Chinese and Jews, have extensive family networks that extend business relationships beyond the usual boundaries. Relationships, however, constrain the amount of capital available to financial growth. To overcome this constraint and increase the scale of business, an economy must supplement (and eventually supplant) relational finance with private finance. In countries where banks dominate finance, an elite group of wealthy insiders lend to businesses based on private information. Bank finance corresponds to the second stage of the development of a start-up in Silicon Valley. As countries become affluent, however, they usually supplement private finance with public finance, which means selling stocks and bonds to the general public. This third stage of finance requires public capital markets such as stock exchanges. Figure 1.4 depicts the three stages of finance in economic development.

Generally, it is the poorest countries that rely on relational finance. This category finances the initial stages of growth in a start-up as in China. As wealth accumulates, however, relational finance constrains further growth. We know of no modern country that became wealthy by relying exclusively on relational finance. Richer countries

supplement relational finance with private finance. The richest countries supplement relational and private finance with public finance. While the richest countries have all three forms of finance, the mixture of private and public finance varies significantly from one country to another. Japan and northern Italy, for instance, have achieved affluence mostly through relational and private finance, with relatively little public finance. The USA (United States of America) and Great Britain, in contrast, have gone onto the third stage and rely mostly on public finance for mature industries now. Germany appears to be shifting away from private finance and towards public finance.

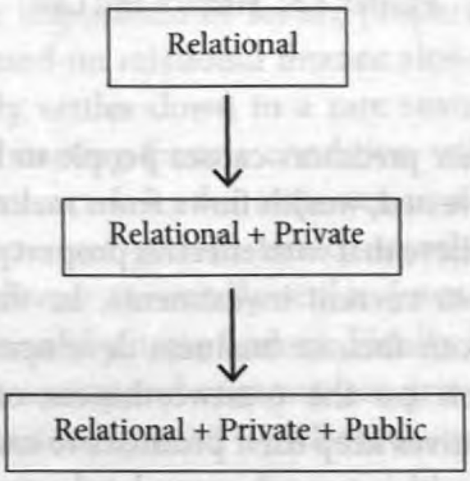


FIGURE 1.4: Finance and Development

Source: Author.

1.5 LAW FOR GROWTH

What determines a country's stage of financial development? We have a simple answer: effective law. Law is effective when legal obligations are enforced sufficiently so that most people keep them. The most relevant laws for economic growth concern property, contracts, and business. Development of effective law in these three areas corresponds to the three levels of finance in Silicon Valley. As depicted in Figure 1.5, relational finance requires property law, private finance requires contract law, and public finance requires business law.

Without effective property protection, people fear that thieves will steal their wealth. Insecurity against hoodlums, mafia, cheating accountants, Ponzi artists, conniving state regulators, corrupt

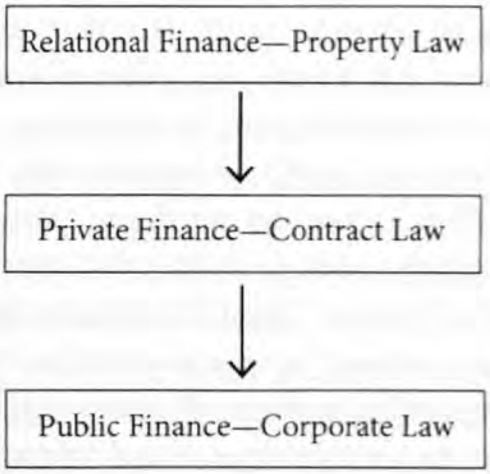


FIGURE 1.5: Finance and Law

Source: Author.

politicians, and other predators causes people to hoard their wealth. Instead of being invested, wealth flows from makers to protectors. In contrast, owners believe that with effective property law, they will enjoy future rewards from current investments. In these circumstances, extended families can finance business development. Relationships provide information on the trustworthiness of others. Informal sanctions make relatives keep their promises to each other. Relational finance can get by without much more legal support from the state than protection of property. However, clans or gangs can protect property, while an effective state is much more reliable.

As development proceeds, sustained growth requires finance from non-relatives. Relationships among non-relatives are too thin for informal mechanisms to carry the burden of enforcing promises. To cooperate together, strangers need to make promises backed by sanctions that third parties enforce. To flourish, private finance requires effective contract law, not just basic property law. Contract law underpins markets for loans, banks, and foreign direct investment. Flourishing private finance also leads to specialized laws for debt collection, bonds, and banking. An effective state enforces promises much more reliably than clans or gangs. Flourishing private finance requires effective contract enforcement.

With private finance, investors retain substantial control over how firms use their investments. When finance moves to the third stage,

the general public buys stock or bonds, and they have little control over the use of their money. Instead, their money comes under the control of the 'insiders' who manage the firm. Insiders have many opportunities to appropriate 'outsiders' investments. For example, they could use accounting tricks to convert profits into salaries, thus depriving stockholders of their dividends. Protecting outsiders from insiders requires more than property and contract law. The additional protection comes from the law of securities, corporations, and bankruptcy, which we call 'business law' or 'corporate law'.

These facts demonstrate how finance constrains innovation and sustainable growth in an economy. Starting from a condition of lawlessness, the imposition of secure property law can cause a spurt of growth based on relational finance alone. The growth rate, however, eventually settles down to a rate sustained by relational finance. Similarly, starting from a condition with secure property and ineffective contract law, the imposition of effective contract law can cause a spurt of growth based on relational and private finance. The growth rate eventually settles down to a rate sustained by relational finance and private finance. Finally, if effective business law supplements property and contract law, growth spurts and then settles down to a rate sustained by *all* three—public, private, and relational finance.

1.6 WHAT MAKES LAW EFFECTIVE?

Growth requires living law, not merely law-on-the-books. When we speak of 'effective law,' we mean rules that control behaviour, not merely written rules. What makes laws effective? Many laws are obligations backed by sanctions. Sanctions give the victim a credible threat against the potential injurer. Foreseeing sanctions, the potential injurer usually does no wrong.[2] Effective law deters wrongdoing and ineffective law does not.

[2] To deter a rational person from doing wrong, the expected sanction should equal or exceed the person's gain from wrongdoing. The expected sanction equals the probability of the sanction times its magnitude. A sanction worth 100 dollars applied with probability ½ will deter wrongdoing by a rational person whose gain does not exceed 50 dollars. Applying the sanction is unnecessary except when the wrongdoer is too irrational to be deterred.

The sanctions that make law effective come from society and not just the state.[3] Early anthropologists decisively proved this fact by studying stateless societies. Instead of speculating about the 'state of nature' from his room in London, Bronislaw Malinowski travelled to the Trobriand Islands in 1914 and observed how the inhabitants resolved their disputes. He found that when one person harmed another, Trobriand Islanders used social pressure to force the injurer's family to compensate the victim's family.[4] Facts like these persuaded anthropologists that law is much older than the state.

Thus, social sanctions remain important in modern societies. This is because they are flexible and cost effective/affordable, so the victims of wrongdoing in business rely on them first. When a businessman breaches a contract, for example, the victim may stop trading with the injurer (refusal to deal), break promises owed to the injurer (retaliatory breach), sully the injurer's reputation (reputational sanctions), and encourage others not to deal with the injurer (boycott).

To improve the efficiency of social sanctions, people use non-state organizations. Thus most uncut diamonds are traded without written contracts in a small number of exchanges in cities like Manhattan and Antwerp. The diamond exchanges have merchant courts to resolve disputes without relying on state sanctions. Banishment from the exchange, which ruins a diamond dealer's livelihood, is the ultimate punishment. [5]

[3] 'Effective law' is therefore a shorthand for several institutions including court law, which protect property, enforce contracts, and regulate businesses. This is similar to the view of Dani Rodrick: 'The cross-national literature has been unable to establish a strong causal link between any particular design feature of institutions and economic growth. We know that growth happens when investors feel secure, but we have no idea what specific institutional blueprints will make them feel more secure in a given context. The literature gives us no hint as to what the right levers are. Institutional function does not uniquely determine institutional form.' See D. Rodrick (2006), 'Goodbye Washington Consensus, Hello Washington Confusion? A Review of The World Bank's Economic Growth in the 1990s: Learning from a Decade of Reform', *JEL*, XLIV (973–87), p. 979.

[4] In the Trobriand Islands, as among other tribal people in Polynesia, liability law traditionally crowded out criminal law. Malinowski's empirical discovery of this fact contracted a priori beliefs of theorists of tribal life. B. Malinowski (1926), *Crime and Custom in Savage Society*, New York, Harcourt Brace & Company Inc.

[5] Lisa Bernstein (1992), 'Opting Out of the Legal System: Extralegal Contractual Relations in the Diamond Industry', *J. Legal Studies*, vol. 21, pp. 115–57.

Instead of decreasing over time, perhaps people will come to rely more on social sanctions in the future. By making information easier to obtain, the Internet has increased the effectiveness of reputational sanctions, especially by posting evaluations from buyers. Reputational sanctions on the Internet are so efficient that strangers buy antiques online without examining them.

Social sanctions, however, are insufficient to prevent wrongdoing in some transaction, especially in one-time transactions with high stakes. Thus car dealers and real estate agents are notoriously exploitative. Even ordinarily moral people can be ruthless when buying or selling a car or house. In big deals, people need the state behind contracts, much like diplomats need an army behind foreign policy. Financing innovation is a big deal with high stakes. Overcoming the double-trust dilemma in private or public financing requires effective state law to prevent stealing by the worst people and cheating by ordinary people.

Figure 1.6 depicts three sources of sanctions that make law effective. Social norms and non-governmental institutions are the first cause of effective law. Social sanctions are applied spontaneously as with gossip or formally as in merchant courts.



FIGURE 1.6: Effective Law

Source: Author.

Second, civil servants in the state bureaucracy apply administrative sanctions, such as revoking permits, cancelling contracts, applying regulations, investigating violations, or imposing fines. Autocratic states especially rely on administrative sanctions to protect

citizens. Thus there have been instances of the state bureaucracy in contemporary China, and the Communist Party that stands behind it, protecting the sources of economic growth by guaranteeing most property rights and enforcing many contracts. To illustrate, imagine that an industrial enterprise and a farming enterprise in Guangzhou dispute the right to a piece of land. They might first appeal to powerful private persons to mediate the dispute. If private mediation fails, they might turn next to a local official in the city government. If one of them rejects the local official's decision, the next appeal might go to a communist party official in Beijing. By this chain, a combination of social and state sanctions deters wrongdoers. Some observers even believe that China has effective property and contract law without having effective courts.

When Europeans look to the state to protect their property and contracts, they ultimately rely on courts. Thus the victim of a broken contract may file a civil complaint against the injurer and threaten to sue for compensatory damages. To be effective, a threat to sue must be credible, which means that the plaintiff stands to gain more in damages from the court than his costs of litigating. Costs of litigating include delays and lawyers' fees. By keeping litigation costs down, courts increase the credibility of threats to litigate by the victims of wrongdoing and enable them to extract settlements on favourable terms. When courts resolve routine business disputes efficiently, most favour the terms set by the party on the right side of the law. In many poor countries, however, inefficient or corrupt courts decrease the credibility of threats to sue and prevent the party in the right from settling on favourable terms.

We propose a 'legal theory' of economic growth. As summarized in Figure 1.5, relational finance requires effective property law, private finance requires effective contract law, and public finance requires effective business law. In our view, the most fundamental defect in law that retards economic growth in poor countries is ineffective property law, followed by ineffective contract law, and ending with ineffective business law. Ineffective laws affect the economy like a toothless gear that spins without moving the vehicle forward.

Writing down a law does not make it effective. Written law in a poor country often resembles written law in a rich country. Thus

property and contract law-on-the-books in India and Nigeria resemble English common law, and property and contract law-on-the-books in Peru resemble the Spanish civil code. The written laws, however, are less effective in India, Nigeria, and Peru than in England or Spain. Citizens enforce legal obligations informally through social pressures, reputation, esteem, and middlemen. Administrators and courts enforce legal obligations through state sanctions. Society and state must work together to enforce law. Securing property, enforcing contracts, and providing a legal foundation for firms requires social norms, state agencies, and courts, not just one of these.

1.7 INSTANCES

The legal progression in Figure 1.5 is from anarchy to property, from property to contracts, and from contracts to business law. Each step increases demands on the legal institutions of society and state. Next we provide some examples of this pattern.

At the lowest level, ineffective protection of property rights devastates an economy, as illustrated by this example from the 1990s.

African Diamonds: Diamond miners in central Africa use hand-tools to dig in a river-bed under the guard of teenage soldiers with Kalashnikov rifles. The miners sell the diamonds to a military officer at a small fraction of world market prices. The diamonds subsequently pass through various intermediaries until they reach Europe. Finally a courier arrives at the central railway station in Antwerp, Belgium, walks quickly to one of the nearby diamond shops; the merchant examines the diamonds and pays in cash, and the courier leaves the city by train within an hour.[6]

In central Africa, producing and transporting diamonds in recent years occurred in conditions that approached anarchy; so central Africa produced fewer diamonds and received much less than the world price for them. If anarchy were replaced by a secure system of property rights, central African nations could produce diamonds with better technology, export them through the regular channels of trade, and receive the world price. (We say nothing here about the unspeakable cruelties and heinous abuse of human rights from this region's anarchy).

Moscow Security: A man opens a small shop selling household goods in Moscow in 1992. A month later three young men visit him with copies of his bank records.

[6] Authors' stories.

Using these numbers, the men calculate a monthly fee that he must pay them to 'protect his shop from hooligans'. If he does not pay, they will destroy his shop. The shopkeeper pays and his business succeeds.[7]

Unlike diamond thieves, Moscow criminals who sell security do not want to take everything from their clients. Selling protection presupposes something to protect. In this example, the Moscow criminals impose a 'security tax' that leaves room for the shopkeeper to succeed. When organized criminals provide security, however, the 'tax' is much higher than when a successful state provides it. Security is a 'natural monopoly,' which means that states can provide it more cheaply and reliably than private parties. The Moscow criminals burden business much more heavily than successful states that provide security. Taken together, the preceding examples illustrate that while private security of property is better than anarchy, it is not better than effective state law.

Now we turn to contracts:

Indian Textiles: In Tamil Nadu in the 1990s, a businessman manufactures cloth, makes the cloth into dresses, hand-embroiders them, and exports the finished product. The entire process occurs inside a single factory where the raw materials, cotton and silk, come in the door only to leave as decorated dresses. Managers in the factory are mostly relatives of the owner. Rural households outside Tamil Nadu would gladly do the hand-embroidery at lower wages than city factory workers. The businessman, however, is unwilling to leave the dresses in rural households in exchange for a promise to embroider them.[8]

In this example, the businessman gathers everyone needed to produce a particular product in a single factory, where his relatives can monitor the other employees. In countries with weak legal institutions, economic cooperation usually involves people with personal ties, especially relatives and friends. Most people, however, do not have enough relatives and friends to achieve the scale of activity required for affluence. Property and contract law lower the cost of monitoring and extend the cooperation to strangers. This facilitates dispersed production, larger organizations, and wider markets. The example of Indian textiles illustrates that weak contract law can keep trade local and organizations small.

[7] Ibid.
[8] Ibid.

Loans are an important and pervasive type of contract, as illustrated by this example:

Mexican Loan: A poor man in Mexico City wants a refrigerator to preserve food for consumption by his family and to sell on the street corner. To buy the refrigerator, the man needs to borrow money. To lend money, the lender needs security against the debtor's default. The court process for repossessing the refrigerator from a defaulting debtor is too slow and unreliable. Instead, the lender requires the borrower to provide information about his family and friends—telephone numbers, addresses, and place of work, etc. If the borrower falls behind in payments, the lender will apply pressure through family and friends, and, if necessary, use their influence to repossess the refrigerator.[9]

This example illustrates a pervasive obstacle to business in poor countries: the impracticality of collecting debts through courts. Thus Mexican courts assess such low interest rates on delays in collecting court judgments that debtors gain by stringing out the legal process. Similarly, there have been reports of judges regularly taking bribes to decide cases in some countries. An Indonesian friend once told us how cases in lower courts in Indonesia often get auctioned. High-cost debt collection dries up the supply of loans to finance business expansion, like the refrigerator for the street vendor.

In this example, however, the parties found a way around debt collection through the courts by relying on family and friends. The lender apparently found a way to use relational finance in circumstances where the borrower and lender have no relationship. In fact, one of Mexico's richest businessmen, Ricardo Salinas, began to build his fortune by figuring out ways like this one to collect debts from poor people who buy consumer durables.

A different kind of financial problem known as the 'soft-budget constraint' exists in banks in countries with a socialist tradition:

Chinese steel: When the government privatized a steel company in northern China, 33 per cent of the shares were sold to the public, who could resell them freely ('tradable' shares). Of the remaining 67 per cent, 47 per cent were allocated to the government and 20 per cent to insiders, who cannot sell them ('non-tradable' shares). After privatization, the steel company kept losing money. Its managers, who had political influence, pressured a state bank to finance its losses by buying its bonds, which were commercially unsound. Thus, 'soft loans' from the government enabled the steel firm to postpone the painful changes necessary to become profitable.[10]

[9] Ibid.
[10] Ibid.

From China to the Czech Republic, partly privatized companies subsist from soft loans given by the government. In the case of China, the voracious appetite for cash crowds out the bonds of profitable companies that are the engine of China's growth. Without access to the bond market, profitable private firms must rely on relational finance. If the government hardened the soft budget constraint, the bond market would finance growth more effectively.

The preceding examples concern property and contract law. Now we turn to an example of business law.

Ecuadorian stocks: A family has a successful business building shrimp farms on coastal mangrove swamps along the southern Gulf of Guayaquil. To grow faster, the business needs to obtain more capital, either by selling stocks or getting a loan. The family knows that shrimp prices could fall in the international market. If the family sells stocks, investors will receive dividends when shrimp prices are high, and nothing when shrimp prices are low. If the family gets a loan, however, the bank must receive periodic payments, regardless of whether shrimp prices are high or low. To reduce its risk, the family wants to sell stock, not get a loan. The family seeks advice from a financial expert in Guayaquil, who says that the small size of the Ecuadorian stock market makes selling stock impractical. In a recent year, Ecuadorians obtained 150 times more capital from loans and sale of bonds than from selling stocks. Since the family cannot sell stock and it regards fixed-interest debt as too risky, it decides to grow more slowly rather than borrow money for faster growth.[11]

When you invest in a company that you do not control, you run the risk that the people who control it will appropriate your investment. The problem is harder to solve for stocks than loans and bonds. Stocks entitle their owners to a share of profits. A company's managers can manipulate reported profits and make dividends disappear. A stock market cannot flourish unless corporate and securities laws effectively protect non-controlling investors. Unlike stocks, bank loans and bonds prescribe an exact repayment schedule that the borrower must meet or go bankrupt. Courts can enforce a prescribed repayment schedule for a loan easier than profit sharing for stocks. Consequently, the credit market can flourish under conditions where the stock market languishes.

When credit markets flourish and stock markets languish, as in Ecuador, bank loans finance most business expansion. As explained, borrowing at fixed interest rates is riskier for an entrepreneur than

[11] Ibid.

borrowing against a share of future profits. A larger stock market would enable entrepreneurs to spread their risk. The skew in financing away from stocks dampens investment in start-ups and slows the pace of innovation.

1.8 ECONOMIC GROWTH, TRUST, AND EFFECTIVE LAW

Sustained economic growth occurs, according to this contribution, when effective law solves the double-trust dilemma. How does this claim square with the facts? Social scientists try to measure the quality of law and government. One such measure, the 'rule of law index,' combines measures of the independence of courts, court delays, effectiveness of the police, and similar indicators of legality into a single number. The horizontal axis in Figure 1.7 indicates the rule of law index for different nations, and the vertical axis indicates their income per capita.[12] Thus each point in Figure 1.7 represents legality and wealth of a different country. Moving to the right, the upward trend of the points indicates that wealth and legality increase together. Every wealthy country has a high score on the rule of law index.

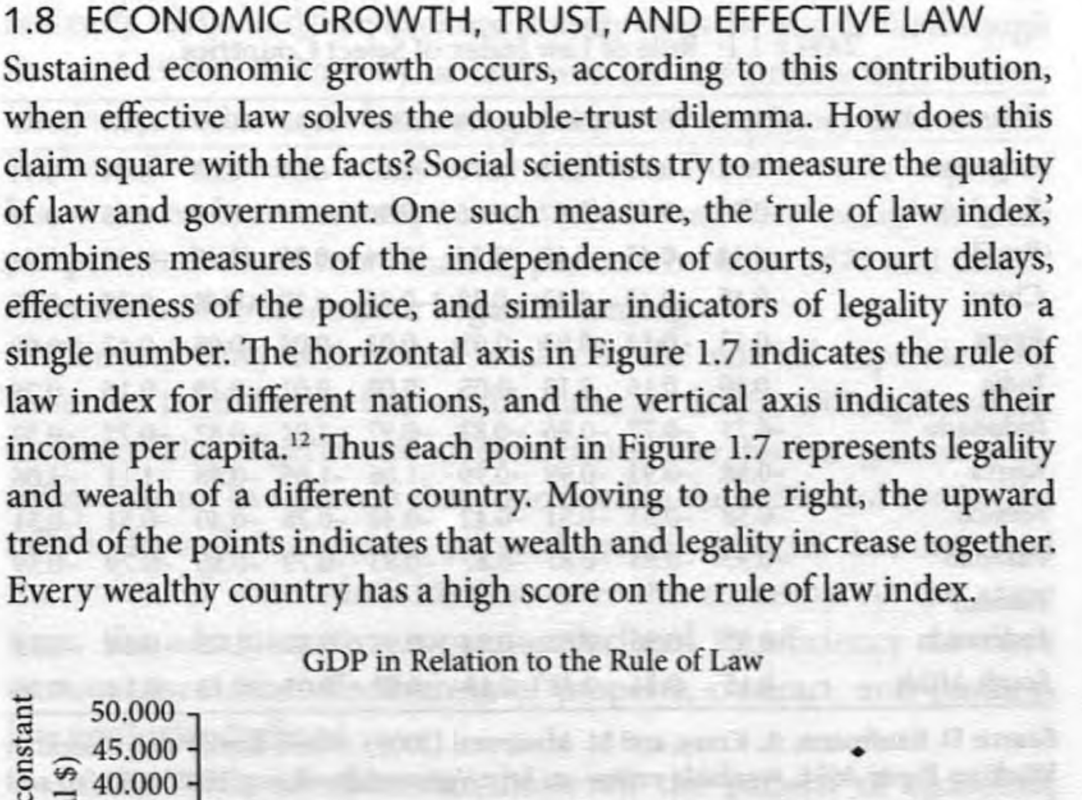


FIGURE 1.7: Per Capita Income and the Rule of Law

Source: World Bank, World Development Indicators, 2004.

Figure 1.7 establishes that wealth correlates with lawfulness, but not the direction of causation. Perhaps wealth causes lawfulness—as they

[12] R. Barro (1997), *Determinants of Economic Growth: A Cross-Country Empirical Study*, Cambridge: MIT Press.

get wealthier, people want more lawfulness just like they want larger houses, better health care, and a cleaner environment. Or perhaps lawfulness causes wealth, which is our hypothesis. It is noteworthy that India has one of the highest values for the rule of law of all bigger developing countries and countries in transition (Table 1.1).

TABLE 1.1: Rule of Law Index of Select Countries

Country / Year	*2007*	*2006*	*2005*	*2004*	*2003*	*2002*	*2000*	*1998*	*1996*
Argentina	−0.52	−0.53	−0.55	−0.73	−0.65	−0.89	−0.05	0.04	0.11
Bangladesh	−0.81	−0.82	−0.87	−0.95	−0.90	−0.79	−0.80	−0.85	−0.77
Brazil	−0.44	−0.45	−0.45	−0.34	−0.34	−0.35	−0.28	−0.28	−0.21
China	−0.45	−0.48	−0.42	−0.38	−0.45	−0.37	−0.44	−0.38	−0.25
Egypt	−0.13	−0.14	0.03	0.03	−0.03	−0.05	−0.05	−0.12	0.08
India	0.10	0.16	0.13	0.05	0.03	0.01	0.19	0.16	0.29
Indonesia	−0.71	−0.77	−0.86	−0.82	−0.97	−1.01	−0.82	−0.77	−0.37
Kenya	−0.98	−0.93	−0.99	−0.99	−1.06	−1.05	−0.98	−1.11	−1.06
Mexico	−0.58	−0.53	−0.51	−0.42	−0.40	−0.38	−0.40	−0.51	−0.51
Pakistan	−0.93	−0.85	−0.87	−0.87	−0.83	−0.79	−0.80	−0.79	−0.59
Russian Federation	−0.97	−0.96	−0.88	−0.84	−0.92	−0.89	−1.05	−0.84	−0.73
South Africa	0.15	0.24	0.18	0.13	0.03	0.05	0.12	0.24	0.26

Source: D. Kaufmann, A. Kraay, and M. Mastruzzi (2008). World Bank Policy Research Working Paper 4654. Available online at: http://go.worldbank.org/2E0SXCR850, and http://earthtrends.wri.org/searchable_db/indexhttp://earthtrends.wri.org/searchable_db/index (last accessed in February 2009). The rule of law index is a number ranging from −2.5 for the lowest value to 2.5 for the highest value of the rule of law.

But notice that the 'rule of law index' captures only that part of the law that we call state law. It consists of information on the perceptions of the incidence of crime, judicial quality and honesty, and the enforceability of contracts by courts and enforcement agencies. It does not capture effective law, which includes social sanctions as well as direct protection by the state.

For now, we test our hypothesis against the world's patterns of growth.

First, consider sub-Saharan Africa, where colonies became independent countries in a process that concluded in the 1960s. After the euphoria of freedom subsided, unresolved ethnic and political conflicts too often devolved into anarchy or civil war. Besides ethnic

strife, the newly independent countries apparently suffered increasing corruption and decreasing competence in administration.[13] These factors caused effective law to decline sharply after independence in the 1960s and 1970s. The situation gradually improved in the 1990s.

Wealth per capita in sub-Saharan Africa increased in the late 1960s and early 1970s, and then declined sharply from the mid 1970s through the mid 1990s. In anarchy or civil war, no one invests their money in a business. Instead of investing, people try to protect their wealth. Without investment, an economy declines. The economic impact of less order and more corruption on Africa was devastating—income per person declined more than 20 per cent between 1975 and 1995.[14] After the mid 1990s, wealth began increasing.

The trajectory of law and wealth in the non-oil dependent oil countries is similar to that of sub-Saharan Africa, although milder. Colonies like Algeria and dependencies like Jordan gained their independence in a process that concluded in the 1950s and 1960s. The fully independent countries pursued socialist policies that increased the power of state administrators over the economy. At the same time, the administrative apparatus declined in efficiency. In these circumstances, the effectiveness of property, contract, and business law gradually declined.

In the Arab oil countries, however, the pattern of economic growth is dependent on world oil prices. The failure of our theory in this case illustrates that economic growth has multiple causes. In technical terms, multivariable simultaneous equations model an economy, not a single variable equation. Besides legality and world prices, other causes of growth include education, population, factor mobilization, regulations, and management of a country's money supply. Our claim is that the effectiveness of property, contracts, and business law matters more than other variables for economic growth per capita in most countries in recent decades (but not in

[13] These are brilliantly depicted in Chinua Achebe's novels, *Things Fall Apart* and *No Longer at Ease*.

[14] GDP per person in sub-Saharan Africa has declined since 1974, roughly by the order of 20 per cent. In 1974 it was 600 US dollars (at constant prices of 2000) per capita. It declined to 470 dollars in 1994 and since then increased slowly to 510 dollars in 2003. See World Bank, World Development Indicators 2005.

all countries in recent years, and not in most countries in past years without peace and free trade).

Now we consider the formerly communist countries of Eastern Europe. When communism triumphed in the first half of the twentieth century, the communist party governed through autocratic state administration. In each country, state planning replaced markets, nationalized industries dwarfed private ownership, and public law crowded out private law. State administration and the communist party guaranteed the stability of property and contract arrangements. This foundation of government and law persisted through the 1960s,1970s, and 1980s, until communism collapsed in 1989. These countries then abandoned planning and tried to introduce a market economy. The change from planning to markets, whose strongest form was called 'shock therapy,' called for strengthening the civil law of property and contracts, and mitigating the scope public law. However, disruption of state administration and the collapse of the communist party removed the two guarantees of stable property and contract arrangements. Immediately after 1989, social norms, the state bureaucracy, and the courts did not effectively protect property or enforce contracts in this period. After 1989, some of the formerly communist countries—Poland, the Czech Republic, Slovakia, Slovenia, Hungary, Latvia, Estonia, and Lithuania—committed to a path resulting in full membership in the European Union in 2004. More recently, Bulgaria and Romania completed a similar process and joined the EU. As a prerequisite for full membership in the EU, these countries radically reformed their legal systems. They had to meet performance targets with respect to eliminating corruption, creating independent courts, and enforcing property and contract rights. The result was a dramatic improvement in the effective law of property, contracts, and business.

The other formerly communist countries—Russia, Belarus, Georgia, Moldova, and Ukraine—did not seek EU membership, and their economic experience was very different. After 1989, the state mostly withdrew its economic controls, but did not reform the institutions of civil and business law in these countries. It is said that 'they got the shock without the therapy'. Instead of getting rich by raising productivity, the best way to prosper in these conditions was to steal

natural resources from the state and sell them abroad. The immediate result was called gangster capitalism. This was accompanied by a deep fall in effective law in 1989, followed by a slowly rising curve.

Former communist countries of Europe follow a similar pattern to their economic growth. The abrupt legal improvements in the countries that joined the EU set off spectacular economic growth. These countries have achieved sustained growth rates well above the rest of the European Union until the financial crisis of 2008. In contrast, economic performance was dismal in the formerly communist countries that did not join the EU. On an average, the per capita income in 2004 had not recovered to its level in 1990. From 1990 to 1998, per capita the income in Russia apparently declined by 42 per cent.[15] It has recovered subsequently, lifted by rising natural resource prices in the world and possibly by more effective law.

Turning to Latin America, almost all countries in the late 1970s and early 1980s ended state ownership of key industries (privatization), reduced the regulation of private business (liberalization), and removed barriers to international trade (free trade). The market was liberalized in a context of weak state protection of investors and competitors. The results resemble a milder form of Russia's transition to markets without institutions. With the remarkable exception of Chile, liberalization and privatization produced worse economic performance in 1990s in Latin America than in earlier decades. When the reform movement faded in the 1990s, more stability returned to finance and property law, and the economy resumed growing.

In China the communist revolution triumphed in 1949 and the country followed the Russian path of replacing markets with planning. The protection of private property and the enforcement of contracts collapsed during China's Cultural Revolution in the 1960s. Starting from the collapse of law in the Cultural Revolution, the reforms under Deng Shao Ping in the 1980s rehabilitated the law of property and contracts, especially by dissolving the agricultural communes and restoring private businesses. Legality improved abruptly in the

[15] This estimate in the World Bank Development Indicators attempts to encompass the illegal, underground economy, which is large and hard to measure. Subsequently, Russia has recovered by becoming more lawful and enjoying rising mineral prices.

1980s and continued improving gradually. Business networks, the state bureaucracy, and the communist party provided protection of property and enforcement of contracts. China replaced state led growth with state protected growth and obtained spectacular results. In contrast, the formerly communist countries of Eastern Europe replaced state-led growth with unprotected property and suffered economic decline. After much progress, however, the level of legality remains low in China (see Table 1.1). To sustain economic growth, China presumably needs to increase legality, which requires better courts and less corruption.

India, is a democratic state with a history of independent courts, good written laws of property and contract, and a plague of court delays and official corruption. After India gained independence from Britain in 1947, state planning gradually crowded out markets, and public law gradually crowded out private law. By 1980 India resembled Russia in having a state-led economy. Subsequently, India reversed these developments, gradually dismantled state planning, and revitalized private law. Instead of shock therapy, India took many small steps towards privatization, liberalization, and free trade. As the state withdrew its economic controls, private property and freedom of contract strengthened. India gradually improved its legal institutions and moved to a market economy, and enjoyed high growth rates for more than 20 years.

In summary, we propose a legal theory of economic growth. Sustained growth requires technological or entrepreneurial innovation in rich as well as in poor countries. Either type of innovation requires solving the double-trust dilemma. The solution is a double bond that comes from effective law of property, contracts, and business. Recent data suggests that changes in the effective law of property, contracts, and business may explain changes in economic growth for many countries. Decolonization in sub-Saharan Africa caused a sharp decline in effective law, which caused a sharp decline in wealth. In the 1980s, many countries replaced state planning with markets. Some of these countries, notably China, India, and the formerly communist countries that joined the European Union, increased the effectiveness of private and business law, which caused economic growth to spurt. Others, notably Latin America (except

Chile) and the formerly communist countries of Europe that did not join the European Union, undermined the effectiveness of private and business law, which caused economic growth to pause or decline. The Arab oil countries do not fit into this model because world oil prices drive their economic performance.

The current data is inadequate for proving the highest generalizations of social science, like oxygen is inadequate for climbing the highest mountains. This data cannot decisively prove or disprove the generalizations in this chapter. (Some of our law school colleagues are uncomfortable with generalizing from imperfect data, although they are comfortable with generalizing from no data at all).

REFERENCES

Rodrick, D. (2006), 'Goodbye Washington Consensus, Hello Washington Confusion? A Review of The World Bank's Economic Growth in the 1990s: Learning from a Decade of Reform', *JEL*, XLIV (973–87), p. 979.

Malinowski, B. (1926), *Crime and Custom in a Savage Society*, New York: Harcourt brace & Company Inc.

Bernstein, Lisa (1992), 'Opting Out of the Legal System: Extralegal Contractual Relations in the Diamond Industry', *Journal of Legal Studies*, vol. 21, pp. 115–57.

Barro, R. (1997), 'Determinants of Economic Growth: A Cross Country Empirical Study', Cambridge: MIT Press.

2

UNIT LINKED INSURANCE POLICIES

Do Regulatory Gaps Pose A Moral Hazard?

T.S. Somashekar

Until 1999 the Indian life insurance sector had a public sector monopoly in the form of Life Insurance Corporation of India (LIC). The Malhotra Committee in its report had recommended liberalization of the insurance sector.[1] The need to increase household's savings to finance infrastructure development finally led to opening up of the industry to private firms in 1999. Further, to regulate the sector, the Insurance Regulator and Development Authority (IRDA) was created in the year 2000.[2] Private companies were allowed to enter the market with a minimum paid-up capital of rupees one billion. Promoters had to have minimum of 26 per cent of the paid-up capital and a maximum of 40 per cent. Foreign capital was restricted at 26 per cent of the equity. Some considered the initial capital requirement as an entry barrier. Nevertheless, by March 2007, there were fifteen firms offering life insurance products and the average annual first year premium growth was a spectacular 47.06 per cent for the period 1999–2007.[3]

[1] For a summary of the report and its recommendations see *The Indian Economic Survey* (1994–5), Ministry of Finance, Government of India, p. 68.

[2] *The Indian Economic Survey* (1994–5), Ministry of Finance, Government of India http://indiabudget.nic.in/es1994–95/3%20Money%20and%20Finance.pdf *The Indian Economic Survey* (1999–2000), Ministry of Finance, Government of India, http://indiabudget.nic.in/es99-2000/chap48.pdf (last accessed on 25 March 2009).

[3] Insurance Regulatory and Development Authority, Annual Report (2006–7), p. 9, http://www.irdaindia.org/ar06-07/ar-2006–07-eng.pdf (last accessed on 28 March 2009).

In the financial year April 2006–March 2007, LIC continued to have a major share of the market, retaining 82 per cent, with the rest going to the predominantly private sector.

Most of this growth in the life insurance sector can be attributed to the development and marketing of Unit Linked Insurance Policies (ULIPs). A ULIP is a life insurance policy that combines risk cover and investment. It has the characteristics of a life insurance and a mutual fund. Along with a life cover it also offers the opportunity to earn market-linked returns. Therefore the risk of the return is transferred onto the investor. Earlier, insurance companies were in the habit of offering endowment policies with guaranteed returns. These were very popular and a substantial chunk of life insurance premium collections were for such policies. But soon, declining interest rates made such guarantees difficult and the IRDA forbade such guarantees. Further, a booming stock market dwarfed the returns guaranteed by insurers.

These circumstances led to the marketing of ULIPs. ULIPs offered the insured an opportunity to invest in equities or bonds or a combination of both, depending on the risk appetite, for better returns. The risk of the return was thus no longer guaranteed, but transferred to the insured. These instruments became so popular that they rivalled mutual funds in terms of attracting household savings and accounted for almost 82 and 46 per cent (of the total first year premiums collected) for private insurers and LIC respectively in 2006–7. But while ULIPs gradually began to resemble mutual funds, their regulations were considerably different. ULIPs investments are regulated by the IRDA while mutual funds are regulated by the Securities and Exchange Board of India (SEBI) (Mutual Funds) Regulations Act, 1996. Differences in these regulations and other tax benefits to insurance products seemed to cause an uneven platform for competing sectors. Besides, the weaker regulatory structure for ULIPs exposed investors to moral hazards. But at the same time, there existed competition between the insurance companies and also between the insurance and the mutual fund sectors. It has been reported that investors in mutual funds and ULIPs are basically the same category of people.[4] Intense competition fills in for initial market

[4] Livemint reports that, 'The Invest India 2007 survey shows that big investors in ULIPs are the same as the ones buying mutual funds and equities.' But large

imperfections that could arise due to information asymmetries or in this case regulatory gaps.

The life insurance sector, with a penetration of 4.10 per cent[5] of the Gross Domestic Product (GDP) has still tremendous potential for growth given the rising incomes and a large middle class. Similarly, the mutual fund industry has also chalked up an impressive growth in assets under systematic management. Both these sectors are bound to be important destinations for household savings, which was around 32 per cent of the GDP during 2006–7. Both sectors have emerged as keen competitors for household savings given their close similarities in term of product profile. This paper aims to examine if the deficiencies in the regulations of ULIP's have in any way rendered lower returns compared to mutual funds.

2.1 REGULATORY DIFFERENCES

Life insurance companies in India are regulated by the IRDA (Protection of Policy Holders' Interests) Regulations, 2002. With this, an archaic 1938 Insurance Act stood amended. Mutual funds on the other hand are regulated by SEBI (Mutual Funds) Regulation Act, 1996. There exists a fundamental difference between the regulations. These could be broadly categorized into issues of transparency, cost, and marketing.

With respect to transparency, mutual funds are required to report their Net Asset Value (NAV) on a daily basis and their detailed portfolio holdings at least on a quarterly basis. Most funds disclose their portfolio on a monthly basis. Such disclosure keeps the investor assured about the risk profile of the fund and prevents deviations from the objective of the fund. In the case of ULIPs, such detailed portfolio disclosure is not forthcoming. It remains unclear as to whether they are required to disclose their portfolio and very few actually do so. This opens the investor to the hazards of dubious investments and improper risk of the portfolio. If the ULIP does indulge in such

investors are more sophisticated than smaller investors who are more numerous. Hence there exists substantial scope for investors to invest in a poorly informed manner. See http://www.livemint.com/2007/07/30234522/India_Unveiled_Ulip-schemes.html.

[5] Insurance penetration is defined as premium volume as a per cent of gross domestic product.

practices then the returns generated could be affected through inferior performance compared to its peers in the mutual fund industry and/or the benchmark index. Further, the NAV of mutual funds is net of all possible expenses of the fund. But for ULIPs the NAV is prior to expenses. These expenses are also quite complex and not uniform across insurers. Investors have no clear idea as to the actual worth of their investment.[6]

The issue of costs is even more perplexing in the case of ULIPs. Equity funds have an overall operating expense cap with a maximum of 2.5 per cent of the assets under management. This decreases as the size of the fund increases. Since these expenses are also exclusive of front-end loads, the total cost could have come to a maximum of 7 per cent of the ULIP. These loads have now been reduced to zero by SEBI when the investment is made directly without the help of an agent. The costs are simple enough to understand for an investor.[7] For a ULIP the front-end cost is phenomenal. It ranges from 25 per cent to 75 per cent. This cannot be justified on the grounds of the additional life cover as most of these go as commissions to agents. Besides, there are other recurring expenses such as mortality charges (for life cover), fund manager charges, and policy and administrative expenses. While the maximum fund management fee is one per cent in the case of equity funds, the usual fee charged is 1.5 per cent for ULIPs. There are no caps on these recurring expenses. ULIPs should normally have lower expenses given that they have a lock-in period similar to close-end funds. It reduces churning in portfolios due to redemptions and thus costs associated with untimely sales and transaction costs. By the same logic, its performance should also be better as investments with a longer horizon can be made compared to open-ended mutual funds which face continuous redemptions and purchases.

[6] A study by McKinsey on Indian Asset Management reveals that consumers view ULIP as a savings or investment instrument and ULIPs have also been marketed as investments. McKinsey (India) (2008), Indian Asset Management: Achieving Broad Based Growth, http://www.mckinsey.com/locations/india/mckinseyonindia/pdf/AMC_Exec_Summ.pdf (last accessed on 12 January 2009).

[7] The McKinsey study (ibid), however, reveals that investors consider mutual funds to be more complex to understand. But this could essentially be due to the diverse types of funds that are now being marketed.

The marketing of ULIPs is driven by high commissions and creates further problems. Insurance companies are allowed to pay 40 per cent of the first year's premium as commission and this reduces to 7.5 and 5 per cent in the second and third year respectively. Since the insurance industry in India is still marketed predominantly via the individual agent route there exists considerable scope for abuse. The insurance industry had more than two million individual agents (as on March 2007)[8] compared to about 57,000 agents for mutual funds.[9] Individual agents themselves did not understand the product and were selling ULIPs like an assured returns product with returns in excess of 25 per cent per annum.[10] The agents had every incentive to aggressively market ULIPs (and not mutual funds) as these offered far higher commissions.[11] Indian household investors are still largely dependent on the advice of these individual agents who have a perverse incentive of over-marketing ULIPs. Such high commissions are not at all justifiable and in no way add value to the insurance component.

Another interesting aspect is that the purchase of mutual funds requires stringent norms for the investor as well. Purchases can be made only via cheque and investors have to disclose the Permanent Account Number (PAN)[12] while making such investments as also other personal details. The objective is to prevent money laundering. There is no such requirement in the case of ULIPs. Purchases can be made with cash and no PAN is needed. This offers insurance companies the opportunity to tap into the large shadow economy of India.

All these factors point to the very advantageous position enjoyed by the life insurance industry compared to the mutual fund sector. It

[8] IRDA Annual Report 2007–8, p. 47. Insurance Regulatory and Development Authority, Annual Report (2007–8) http://www.irdaindia.org/annual_report/annual_rep_2007_08.pdf (last accessed on 25 March 2009).

[9] *Association of Mutual Fund (AMFI) Quarterly*, January–March 2007, page 4.

[10] This author has himself had the experience of being offered a ULIP product with an assured return of 25 per cent and the agent had poor understanding of the returns being linked to the market. Only 20 per cent of the agents are apparently trained as per the IRDA annual report 2007.

[11] The McKinsey survey finds that 80 per cent of the AMFI registered mutual fund agents sold insurance whereas only 55 per cent sold mutual funds. This clearly indicates the hazard of differential commissions.

[12] The Permanent Account Number (PAN) is a ten-digit alphanumeric number issued by the income tax department for all people submitting their tax returns.

has certainly contributed to the faster growth of life insurance. But the crucial question is how this has affected the buyers of ULIPs.

2.2 LITERATURE REVIEW

The costs and benefits of portfolio disclosure are well documented in literature. David K Musto[13] finds that agency problems are related to portfolio disclosures. Funds normally indulge in window dressing of portfolio before the disclosure date in order to get rid of riskier assets which they held before. The less frequent the disclosure the easier it would be for a fund to do this. Weili Ge and Lu Zheng[14] find that American funds with less frequent disclosures are like to have higher portfolio churning, expenses and also the likelihood of committing fraud. At the same time, Russ Wermers[15] points out that more frequent portfolio disclosures can also have a negative impact on returns due to higher costs of compliance and front running.[16]

Besides moral hazards, the large size of the assets of the ULIPs can themselves erode returns. Chen et al.[17] showed how pre- and post-expense returns declined with size for US funds due to liquidity issues. Similarly, Somashekar[18] has shown how large Indian funds (asset size greater than Rs 7 billion) performance when controlled for various fund characteristics have tended to underperform smaller funds in

[13] David K. Musto (1997), 'Portfolio Disclosures and Year-End Price Shifts', *Journal of Finance*, 52 (4), pp. 1563–88.

[14] Weili Ge and Lu Zheng (2006), 'The Frequency of Mutual Fund Portfolio disclosure', SSRN No. 557186.

[15] Russ Wermers (2001), 'The Potential Effects of More Frequent Portfolio Disclosure on Mutual Fund Performance', *Investment Company Institute Perspective*, pp. 1–12, http://www.ici.org/pdf/per07-03.pdf (last accessed on 22 December 2008).

[16] Professional traders would be able to guess trades of funds and trade ahead, thus putting the funds at a distinct price disadvantage. This affects the future particularly as they know what stocks would be sold to provide liquidity to share holders and also gain tax benefits. For more see http://www.ici.org/perspective/per07-03.pdf (last accessed on 22 December 2008).

[17] J. Chen, Hong H., Huang M., Kubik J.D. (2004), 'Does Fund Size Erode Mutual Fund Performance? The Role of Liquidity and Organisation', *The American Economic Review*, 94(5), pp. 1276–1302.

[18] T.S. Somashekar (2008), 'Fund Size and Its Impact on Fund Returns—An Empirical Study of Indian Mutual Funds', *Indian Economic Journal*, 56(1), pp. 102–111.

bearish markets. In a period dominated by more bearish months they are likely to perform poorly compared to the smaller funds. For the many ULIPs in India their pure equity portfolio is substantially larger than Rs 10 billion and it is reasonable to expect a negative impact on their performance.

2.3 METHODOLOGY AND DATA

In order to compare the performance of ULIPs and mutual funds and establish if regulatory differences contributed in any way, it is necessary to control all other factors that possibly influence performance. ULIPs have different investment strategies. The aggressive strategy which had almost 100 per cent investment in equities or growth focus was chosen. All the ULIPs chosen tend to invest in large cap growth oriented stocks. So the sample had to focus on only such ULIP products. Further, since they are of recent origin in India, data availability on NAV was severely limited. In order to maximize the number of ULIPs and the months for coverage, the time period of February 2005 to January 2008 was chosen. The sample finally consisted of seven ULIPs.

To compare likes it was imperative that the mutual funds chosen were similar in characteristics. Large equity diversified funds with asset size persistently above Rs 4 billion were selected. From these only funds with a large growth stock focus were chosen. Finally the sample boiled down to seven mutual funds. This gives a total of 504 months for funds and ULIPs together.

Hence, a balanced panel data is used for this particular study.

The NAV of each is calculated as follows:

$$\text{Returns} = \frac{NAV_t - NAV_{t-1}}{NAV_{t-1}} \times 100 \qquad (2.4.1)$$

NAV_t: end-of-the-month net asset value (NAV)
NAV_{t-1}: beginning of the month

Similarly, monthly returns are computed for the chosen benchmark index—BSE 100.[19] Data on NAV for mutual funds is from AMFI and for ULIPs is from the individual insurance companies.

[19] BSE 100, also popularly termed as the sensex, is an index of the Mumbai (previously Bombay) Stock Exchange, which consists of 100 of the largest capitalized stocks.

In order to compare performance, the risk-weighted returns are calculated using the Sharpe Ratio–S(p). It is calculated as:

$$S(p) = (R_p - R_f) / \text{Std.Dev}(p) \qquad (2.4.2)$$

p is the fund or ULIP
R_p is the average annual rate of return of p.
R_f is the rate of return of the 91 day t-bill.
StdDev(p) is the standard deviation of r_p.

TABLE 2.1: Comparison of the Mean Returns, Standard Deviation, and Risk-weighted Returns of the Sampled ULIPs and Mutual Funds

Name	*Type*	*Mean return*	*Std. Dev*	*Sharpe ratio*
HDFC equity fund	ED	3.25	6.46	0.503096
Birla SL equity fund	ED	3.54	7.27	0.486933
HDFC	ULIP	3.09	6.89	0.448476
HDFC CS	ED	3.1	6.99	0.443491
HSBS equity fund	ED	3.08	6.96	0.442529
Reliance	ULIP	3.03	6.92	0.437861
Kotak Mahindara	ULIP	3.11	7.11	0.437412
Birla SL Basic Ind.	ED	3.28	7.79	0.421053
Birla advantage fund	ED	2.87	6.82	0.420821
ICICI	ULIP	2.64	6.31	0.418384
BSE 100	Index	3.03	7.26	0.417355
HSBC Opp.	ED	3.13	7.51	0.416778
Birla Sunlife	ULIP	3.06	7.68	0.398438
Bajaj Allianz	ULIP	2.7	6.85	0.394161
Metlife	ULIP	2.44	7.08	0.344633

Source: Author.

Table 2.1 shows the average returns of the sample along with that of the index. The Sharpe ratio–S(p)—is the portfolio returns in excess of the risk-free rate of return (in this case the 91 day treasury bill) divided by the portfolio standard deviation. A large number of the sampled ULIPs have underperformed on the index in terms of the risk-weighted returns as given by the Sharpe Ratio.[20] Mutual funds

[20] The Sharpe ratio is the ratio of the portfolio excess returns over a risk free rate of return divided by the standard deviation of the portfolio. W.F. Sharpe (1966), 'Mutual Fund Performance', *Journal of Business*, vol. 39, pp. 119–38.

have done better with only one slipping below the index. However this still does not give us a clear picture as to whether the difference in performance and stock picking ability is statistically significantly different. Hence a single factor capital asset pricing model is used to asses the same. The model used is specified as below:

$$R_p - R_f = \alpha + \beta\,(R_{BSE100} - R_f) + D + e \qquad (2.4.3)$$

R_{BSE100} – monthly return of the index.
D is a dummy variable which is 1 if it is a ULIP and 0 if a fund. A significant negative value for the dummy would mean inferior-stock picking talent on the part of ULIPs.

Beta here represents the systematic risk due to general market conditions that cannot be avoided by diversification of the portfolio. It is the measure of a fund's risk in relation to the market. A beta of 0.8 means the fund's total return is likely to move up or down 80 per cent of the market change. A beta of 1.2 means total return is likely to move up or down 20 per cent more than the market.

Alpha (α) represents the return generated based on idiosyncratic risk. It represents the return due to deviation from the market risk. If positive, it represents stock selectivity skills resulting in superior performance. If negative, it represents inferior stock selection skills. A lower average alpha for ULIPs would also strengthen the case that there exists some degree of moral hazard in stock selection.

2.4 EMPIRICAL RESULTS

TABLE 2.2: Relative Performance of Open-ended Equity Funds and ULIPs without Excluding Expenses of ULIPs

Variable	*Coefficient*	*Std error*	*T-statistic*	*Prob*
Alpha	0.40	0.18	2.15	0.03
Beta	0.90	0.02	51.18	0.00
D	–0.36	0.25	–1.45	0.14

Source: Author.
Note: R squared 0.84.

The regression results are on predicted lines. The model explains 84 per cent of the NAV trends. The beta for the combined sample is

0.9, showing a divergence from the risk pattern of the index. The alpha is positive at 0.39 and is highly significant. This indicates positive returns due to superior stock-picking talent. The crucial variable is the alpha. The alpha has a negative sign which means that the insurance products have inferior stock-picking talent. But it is not significant at the 10 per cent levels. This means that the stock-picking ability of the insurance portfolio managers are on an average the same as that of the mutual fund managers.

This is a crucial finding, as the regulatory gap should have led to inferior stock selection given the information asymmetry and the consequent potential for moral hazards. But apparently competitive forces have filled in for the regulatory gap. This blunts the criticism against ULIPs.

However, once expenses are accounted for, the returns of the insurance products are lesser by at least 1.5–2.5 per cent per annum. In monthly terms it would be lesser by 12.5 to 21 basis points. The same model is again applied for the panel data after deducting an average 0.17 per cent per month for the ULIPs.

TABLE 2.3: Relative Performance of Open-ended Equity Funds and ULIPs Excluding Expenses of ULIPs

Variable	*Coefficient*	*Std error*	*T-statistic*	*Prob*
Alpha	0.40	0.18	2.15	0.03
Beta	0.90	0.02	51.18	0.00
D	–0.53	0.25	–2.13	0.03

Source: Author.
Note: R squared 0.84.

In this case we find that all the other variables including the alpha and the beta retain the same values but the dummy variable capturing the possible inferior stock selection changes. It is (–)0.53 and is significant at 5 per cent levels. Subtracting the dummy value from the positive alpha gives us an overall alpha of (–)0.14 for the insurance products. Hence when comparing the genuine net performance which is not revealed in the NAV of the ULIPs, we find an inferior performance compared to the mutual funds.

This is an issue for concern. Most of the insurance companies selected also have mutual funds as can be seen in Table 2.1. It raises

the question as to why the stock selection of their insurance portfolios is inferior to the open-ended equity funds. While a simple poor selectivity might not indicate a moral hazard, it definitely strengthens the case when their own equity funds seem to do better.

2.5 THE NEED FOR IMPROVEMENT IN ULIPS

The insurance sector in India has grown considerably since its deregulation. There have been innovations in terms of products. The regulatory structure, is found wanting. It, however, has a close cousin in terms of SEBI (Mutual Fund) Regulations for learning how to regulate ULIPs that are significantly similar to mutual funds. This does not seem to have happened. Its weaker set of regulations, both in terms of disclosure and in terms of containing expenses, has led to a distinct possibility of inferior stock selectivity. Investors' choice of the insurance route for investment is led by individual agents with a perverse incentive and based on poor information caused by regulatory lags. The huge agent commissions allowed in the insurance sector may also divert investments from the more efficient equity mutual funds to less efficient ULIPs. In an efficient market one would expect resources to flow to the most productive user. The combination of insurance is no excuse for poor performance. The IRDA would do well to improve disclosure of portfolio and also regulate commissions and expenses. Perhaps the situation may also warrant a 'super-regulator' with financial products becoming increasingly less differentiated.

REFERENCES

Insurance Regulatory and Development Authority, Annual Report (2006–7) http://www.irdaindia.org/ar06-07/ar-2006-07-eng.pdf (last accessed on 17 December 2009).

Insurance Regulatory and Development Authority, Annual Report (2007–8) http://www.irdaindia.org/annual_report/annual_rep_2007_08.pdf (last accessed on 17 December 2009).

J. Chen, H. Hong, M. Huang, J.D. Kubik (2004), 'Does Fund Size Erode Mutual Fund Performance? The Role of Liquidity and Organisation', *The American Economic Review*, 94(5), pp. 1276–1302.

M.C. Jensen (1968), 'The Performance of Mutual Funds in the Period 1945–1964', *Journal of Finance*, vol. 23, pp. 389–416.

Weili Ge and Lu Zheng (2006), 'The Frequency of Mutual Fund Portfolio disclosure', SSRN No. 557186.

J. Hirshleifer (1975), 'Speculation and Equilibrium: Information, Risk, and Markets', *Quarterly Journal of Economics*, vol. 89, pp. 519–42.

McKinsey (India) (2008), *Indian Asset Management: Achieving Broad Based Growth*, http://www.mckinsey.com/locations/india/mckinseyonindia/pdf/AMC_Exec_Summ.pdf (last accessed on 17 December 2009).

W.F. Sharpe (1966), 'Mutual Fund Performance', *Journal of Business*, vol. 39, pp. 119–38.

T.S. Somashekar (2008), 'Fund Size and Its Impact on Fund Returns—An Empirical Study of Indian Mutual Funds', *Indian Economic Journal*, 56(1), pp. 102–11.

The Indian Economic Survey (1994–5), Ministry of Finance, Government of India http://indiabudget.nic.in/es1994-95/3%20Money%20and%20Finance.pdf (last accessed on 17 December 2009).

The Indian Economic Survey (1999–2000), Ministry of Finance, Government of India, http://indiabudget.nic.in/es99-2000/chap48.pdf (last accessed on 17 December 2009).

Russ Wermers (2000), 'Mutual Fund Performance: An Empirical Decomposition into Stock-Picking Talent, Style, Transaction Costs, and Expenses', *Journal of Finance*, vol. 55, pp. 1655–1703.

—— (2001), 'The Potential Effects of More Frequent Portfolio Disclosure on Mutual Fund Performance', *Investment Company Institute Perspective*, pp. 1–12, http://www.ici.org/pdf/per07-03.pdf (last accessed on 17 December 2009).

3

REPAYMENT BEHAVIOUR UNDER JOINT LIABILITY LOAN CONTRACTS IN A COMPETITIVE ENVIRONMENT

Alexander S. Kritikos*

3.1 MICROFINANCE INSTITUTIONS IN INDIA

In the past decade thousands of Microfinance Institutions (MFIs) mushroomed out of every corner in Asian countries like India and Bangladesh, in developing countries of Africa and South America, and in the transition economies of Eastern Europe. As the number of MFIs increased, some pioneers of the Microfinance approach started to complain that the sector had turned into a multi-billion-dollar industry that is destroying the social component of the initial movement. Many MFIs formerly acting in a kind of regional monopoly—being in competition only with the local money-lenders—started competing for clients. In India, for instance, competition substantially increased in the southern states, in areas around Guntur, the West Godavari region of coastal Andhra Pradesh, and the Thiruchirapalli–Madurai region of Tamil Nadu.[1] From the clients' perspective competition seems of course to be a positive development in a market segment originally started by non-profit-organizations. These are usually guided by targets other than profit maximization, for instance providing poor

* I am very grateful to Denitsa Vigenina for the research assistance. This paper was presented in 2008 at the conference on the 'Economic Analysis of Law in India' in Hyderabad, India.

[1] Microfinance Information Exchange (MIX) (2007), *India Microfinance Review*, Washington.

people with access to credit. Competition usually drives prices down and enhances products, even when these are served by NGOs. In this case the market of microfinance is expected to improve in terms of efficiency and product quality.

However, it is also a well-known fact that in credit markets information asymmetries may produce inefficient outcomes if competition is unregulated. The most obvious problem in an unregulated environment is that MFIs may be confronted with a decline in their repayment performance if their clients are able to simultaneously borrow from second sources.[2] McIntosh *et al.*[3] founded that in Uganda, rising competition between the MFIs 'lead to a gradual deterioration in repayment performance and to a drop in savings,' because the MFIs had 'no formal information sharing mechanism about clients' credit histories.' They conclude that institutional innovations have to be developed for information sharing on client indebtedness levels if the major negative impact of increasing competition—namely declining repayment rates—should be avoided.

The majority of MFIs in India uses some sort of joint-liability loan contracts similar to the so called 'Grameen Bank approach.' There seems to be a kind of common sense among many practitioners of microfinance that joint-liability contracts do better than individual lending schemes when an MFI is facing increasing competition. Since the MFI's screening, monitoring, and enforcement cost of clients ought to be smaller under joint-liability—part of these duties are transferred under these contracts to the borrowers—MFI practitioners expect to be better able to offer the lowest interest rates if they use joint-liability instead of individual loan contracts.

As was shown in several theoretical approaches (for a thorough overview, cf. Morduch)[4], MFIs using joint-liability loan contracts may look ahead to high repayment rates: First, joint liability sets free strong incentives among the borrowers to self-select into homogenous

[2] K. Hoff and J. Stiglitz (1990), 'Imperfect Information and Rural Credit Markets—Puzzles and Policy Perspectives, *The World Bank Economic Review*, 4, pp. 235–50.

[3] C. McIntosh, A. Janvry, and E. Sadoulet (2004), 'How Rising Competition Among Microfinance Institutions Affects Incumbent Lenders,' Disc. Paper, San Diego.

[4] J. Morduch (1999), 'The Microfinance Promise,' *Journal of Economic Literature*, vol. 37, pp. 1569–1614.

groups of low risks before they sign the loan contracts. Second, if, after the loan disbursement, any group member intends to reject loan repayment, joint liability induces different kinds of peer pressures within the group, mostly with the consequence that all loans are repaid on time even if some group members intended to reject compliance.

Recent empirical research indicated that the behaviour of credit groups who signed joint-liability loan contracts may meet these expectations raised by theoretical literature. However, these results were found under the condition that the respective MFI does not face client competition. A central mark of joint liability lies in the advantage that the effective fees (interest rates plus the cost of joint liability) are lower for groups of low risk borrowers than for groups of high-risk borrowers. This is because the probability of the former group defaulting in making repayments is lesser. Because of this cost advantage for groups of low risks, MFIs may expect that only groups consisting of low-risk borrowers to apply for a joint-liability loan.

In the present paper, we will now empirically analyse the impact of competition on groups of borrowers who (co-)signed joint-liability approaches. We aim to explore whether competition reduces the power of the self-selection process; whether the peer pressures within the group of borrowers are different in an environment of competing MFIs; and to what extent changes in the intra-group behaviour have an impact on the repayment performance or the total lending cost of the borrowers. In order to answer these questions, we developed a questionnaire which was used in two different MFIs with very similar joint liability approaches. As it was not possible to get access to such data in India, we decided to make this analysis in transition economies of the former Soviet Union where the Grameen Bank approach was applied and where the access to data was possible in a way that the aimed comparative analysis could be conducted. The first MFI is Constanta situated in Georgia where—at least during our survey—there was only very low and indirect competition between this MFI and a second MFI named Microfinance Bank Georgia which was offering only individual loans with higher loan volumes.[5] The

[5] For details see D. Vigenina and A.S. Kritikos (2004), 'The Individual Micro-Lending Contract: Is It a Better Design than Joint-Liability?' *Economic Systems*, vol. 28, pp. 155–76.

second MFI we surveyed was FORA in Russia which—in contrast to Constanta—faced direct competition for potential or existing clients.

The rest of the paper is organized as follows. In section 3.2 we provide a short overview of the theoretical and empirical literature on group lending to the extent that is relevant for the present research. In Section 3.3 we describe the group lending methodologies used by the MFIs Constanta and FORA and compare the two approaches with respect to the question whether FORA—working in a competitive environment—is offering 'better' or 'cheaper' products than Constanta. In Section 3.4, the empirical results are presented, showing that in FORA the groups are randomly put together while in Constanta borrowers are self-selected in groups of homogenous risks. We further show that the clients of FORA compensate the missing self-selection by a more intense and effective monitoring process and by a higher willingness of the group to repay the loans of delinquent partners. Finally, Section 3.5 discusses whether the main advantage of competition between MFIs might be overcompensated by higher indirect cost which may specifically arise when using joint-liability contracts.

3.2 APPROACHES IN THE LITERATURE

Theoretical literature provides a comprehensive overview about the mechanism of joint liability, in particular how it alleviates the difficulties of adverse selection, moral hazard, and repayment enforcement, problems which have to be solved in any loan contract. We may differentiate between three stages during the process of employing joint-liability contracts, where this contract design might influence borrowers' behaviour. At the first stage, borrowers have to form groups before they sign the contract. After contracts have been signed and loans are disbursed, borrowers may choose at the second stage between several peer activities, and at the third stage they jointly have to decide about the repayments of the loans.

Thus, starting backwards, at the final stage borrowers are confronted with joint liability in combination with the termination threat (as a sanction) if they fail to repay all loans of a group and they receive the perspective to get access to higher loan volumes (as a positive incentive) if they repay on time. Under joint liability the

whole credit group is considered as being in default, losing access to subsequent loans if even one borrower rejects to repay a small share of his loan. The other borrowers of the group have then to decide whether they will repay for their delinquent partner. If they need further and larger loans and if the utility of getting these loans from the same MFI is larger than the cost of (temporarily) repaying the loan of a delinquent group member in addition to their own loan, these other borrowers of the same group will make the additional loan repayment for the delinquent borrower.[6]

Anticipating the possibility of increased repayment costs (or alternatively of potential exclusion from access to further loans) might induce borrowers to start acting already at the previous (the second) stage of this repayment process: In order to avoid the probability of exclusion from further loans, or the necessity to repay for a delinquent peer at a minimum, borrowers might start to either monitor the partners of their group, put pressure on their peers, or to support them.[7]

Applying peer monitoring means that borrowers make sure that each borrower undertakes investments leading to income streams that enable loan repayments with maximum probability. Under peer pressure, it is understood that the group of borrowers exerts social pressure on delinquent borrowers to repay their loans if they reject repayment in the first place. Peer pressure is expected to be particularly used if the rest of the group has the information that a delinquent

[6] T. Besley and S. Coate (1995), 'Group Lending, Repayment, Incentives and Social Collateral, *Journal of Development Economics*, vol. 6, pp. 27–50. B. Armendariz de Aghion (1999), 'On the Design of a Credit Agreement with Peer Monitoring', *Journal of Development Economics*, vol. 60, pp. 79–104. See also A.S. Kritikos (1999), 'Micro-credits with Incentive-Guided Solidarity Action', *Kredit and Kapital*, vol. 32, pp. 393–425.

[7] J.E. Stiglitz (1990), 'Peer Monitoring and Credit Markets', *The World Bank Economic Review*, vol. 4, pp. 351–66. H.R. Varian (1990), 'Monitoring Agents with Other Agents', *Journal of Institutional and Theoretical Economics*, vol. 146, pp. 153–74. A. Banerjee, T. Besley, and T. Guinnane (1994), 'Thy Neighbor's Keeper: The Design of a Credit Cooperative', *Quarterly Journal of Economics*, vol. 109, pp. 491–515. Besley and Coate (1995). D. Hulme and P. Mosley (1996), *Finance Against Poverty*, London: Routledge. Armendariz de Aghion (1999). Kritikos (1999). Morduch (1999).

borrower is responsible for the repayment problems. In contrast to this, peer support is expected to be applied if a borrower faces repayment problems which were induced through external factors. Peer support might have different meanings like financial support, support at the working place, transfer of information, etc. Thus, with respect to the second stage of this 'repayment game' we may conclude that borrowers of a group might increase efforts like monitoring or supporting their peers or putting them under pressure to repay their loans on time in order to decrease the probability of being made responsible for the repayment of their peers at the final stage.

Since borrowers are also interested in keeping the efforts for these peer activities at a minimum, the joint-liability approach unfolds a further incentive, namely on the matching of the borrowers before they sign contracts. In several theoretical papers it has been shown that similar risk types may group together[8] if they possess sufficient information to recognize each other's risk types.[9] As a result of the matching procedure safe types are expected to team with safe types, and risky types are left with other risky types. More importantly, these approaches explain that a low risk type will reject a matching with a high risk type. The crucial point of this kind of assortative matching is that low-risk borrowers, as they have to make repayments for delinquent peers with lower probability, face considerably lower borrowing costs when they match with another low-risk borrower instead of a high risk type.

As a consequence, if the MFI has sufficiently designed its incentive system (interest payments and joint liability), we may observe the following outcome: 1) High risk borrowers do not apply for this loan type, because the borrowing costs are too high for them if they

[8] See for example, Varian (1990). M. Ghatak (1999), 'Group Lending, Local Information and Peer Selection', *Journal of Development Economics*, vol. 60, pp. 27–50. Kritikos (1999). Morduch (1999). E. Van Tassel (1999), 'Group Lending Under Asymmetric Information', *Journal of Development Economics*, vol. 60, pp. 3–25. J.J. Laffont, and T. N'Guessan (2000), 'Group Lending with Adverse Selection', *European Economic Review*, vol. 44, pp. 773–84.

[9] See Hoff and Stiglitz (1990). S. Khandker (1998), *Fighting Poverty with Microcredit: Experience in Bangladesh*, New York: Oxford University Press for the World Bank.

are only able to match with other high risk borrowers; 2) Low risk borrowers face lower borrowing costs if they realize a homogenous matching with respect to their risk types; 3) The MFI can be almost certain that it is lending its financial capital only to low risk types and has thus solved the adverse selection problem.

This short review of theoretical approaches helps us conclude that joint-liability contracts create three stages of potential decision-making processes where the joint liability contract might unfold activities on all three stages, depending on the ex ante information of the borrowers about their future partners.

A number of empirical studies provide various tests on these theoretical approaches just presented. These studies illustrate that the key components of the group-lending mechanism do not have an unambiguous impact on the borrowers' repayment performance. Their influence varies from country to country probably depending on how the joint-liability contract was designed and processed, and also on the local conditions and cultural peculiarities. With respect to the question on assortative matching, Sadoulet and Carpenter[10] report in their surveys of certain MFIs in Guatemala and Eritrea that most borrowers group randomly in these MFIs and use group lending mainly as a form of insurance. In contrast to this, Wenner[11] and Sharma and Zeller[12] find some indirect hints in their surveys of Costa Rica and Bangladesh for assortative matching. Kritikos and Vigenina[13] were the first to discover that in an MFI in Georgia assortative matching with respect to the repayment risk is existent and had a significant impact on the repayment behaviour of the groups where the groups with the lowest risks had fewer problems with delinquency than groups with relatively higher risks.

[10] L. Sadoulet and S. Carpenter (1999), 'Risk-Matching in Credit Groups: Evidence from Guatemala', *ECARES, Free University of Brussels.*

[11] M.D. Wenner (1995), 'Group Credit: A Means of Improve Information Transfer and Loan Repayment Performance', *Journal of Development Studies*, vol. 32, pp. 263–81.

[12] M. Sharma and M. Zeller (1997), 'Repayment Performance in Group-Based Credit Programs in Bangladesh: An Empirical Analysis', *World Development*, vol. 25.

[13] A.S. Kritikos and D. Vigenina (2005), 'Key Factors of Joint-Liability Loan Contracts', *Kyklos*, vol. 58, pp. 213–38.

For the behaviour of borrowers after the loan disbursement, for instance Wydick[14] and Kritikos and Vigenina[15] analysed the extent to which peer pressures were applied within the groups of borrowers. Wydick[16] found that the intensity with which the borrowers employed various internal rules within their groups strongly depends on the environment: groups in urban areas enforce prompt repayment via peer monitoring, whereas in rural areas potential repayment problems are reduced mainly through the willingness to exert social pressure on delinquent peers. Peer pressure had also in the empirical analysis of Kritikos and Vigenina[17] a significant impact on the repayment behaviour within the groups, while peer monitoring was observed only ex post, when a borrower had run into a repayment problem. More importantly, borrowers of that MFI were willing to provide peer support, in particular financial support, when the group had experienced together more than three loan terms.

This overview of the empirical research reveals that it is possible to observe borrowers of some MFIs who behave, to a certain extent, as theoretically predicted. The main aim of the present paper is therefore to study the impact of those components of the joint-liability contract that proved to be crucial in previous research. We make use of the data of two MFIs which employ the same incentive structures and work in similar cultural environments. The main difference between the two MFIs we chose is that one MFI (FORA in Russia) is exposed to direct competition and the second one (Constanta in Georgia, a former member of the Soviet Republic) is not. By comparing the borrowers' behaviour in these two MFIs, this approach allows us to identify what factors might work in the same way (irrespective of the competitive environment) and on what factors of the joint liability approach competition might have an impact.

More specifically, in a first step, we will analyse the screening process between prospective clients, in particular whether groups

[14] B. Wydick (1999), 'Can Social Cohesion be Harnessed to Repair Market Failure? Evidence from Group Lending in Guatemala', *Economic Journal*, vol. 109, pp. 463–75.

[15] Kritikos and Vigenina (2005).

[16] Wydick (1999) and ibid.

[17] Ibid.

match homogeneously with respect to the repayment risk or whether there is random matching. Since we found that there is incidence for assortative matching not at FORA and the FMI Constanta, we will also be able to analyse the extent to which assortative matching is a necessary condition for the success of joint-liability contracts. In a second step after the loan disbursement, we study whether the differing self-selection processes affect the use of peer pressures.

Before we present the sampling design in Section 3.4, we will provide in Section 3.3 an overview the group lending mechanisms employed by the two MFIs to the extent that is relevant for the present study.

3.3 DESCRIPTION OF THE LENDING TECHNOLOGIES OF FORA AND CONSTANTA

The Lending Methodology: The two MFIs studied—FORA (Russia) and Constanta (Georgia)—are NGOs founded by international development organizations in the late nineties. Their mission is to provide quality loan services to entrepreneurs who are striving to develop their businesses but lack collateral. The majority of their clients are traders who need to increase the working capital in order to maintain or expand their business. Both MFIs have a high share of female clients, who constitute around 70 to 80 per cent of the portfolios. The applied lending technology consists of gradually increasing repeated loans given to solidarity groups, thus being very similar to the approaches used in India. Group members guarantee each others' loans, and the entire group debt must be repaid in order to have access to subsequent loans. Loans are disbursed and collected simultaneously for all group members.

In Constanta, the loans have a maximum term of four months and are required to be repaid in weekly instalments. The group size varies from four to fifteen borrowers. The initial loan size is between 50 and 150 Euro. Loans may increase in each cycle by not more than 50 per cent if the previous loan amounts were repaid on time. The MFI charges an interest rate of 4 per cent per month and an administrative fee of 1 per cent of the disbursed amount. The inflation rate at the relevant year was about 4 per cent per year. In FORA, the groups are little smaller, consisting of three to nine borrowers. Here, the loan terms and the repayment schedules are flexible, varying with

the borrowers' needs and the income stream of their businesses. The initial loan size is on an average about 200 Euro. The interest rate is calculated depending on the type of the loan, the loan cycle, the size of the group, and on the region in which the branch is situated, and is fixed at a monthly rate of 3 per cent to 5 per cent. As a rule, the price of the loan depends on the loan cycle and, the number of the group members. In addition, if a client attracts a new borrower in the group, the interest rate on his loan is decreased. The inflation rate at the relevant year was 13 per cent.

The Lending Procedure: With one small but important exception the lending procedure of the two MFIs is identical. Potential clients (interested in getting access to loans by making use of group contracts) are asked by the loan officers of the MFIs to freely self-select in credit groups. The loan officers themselves do not undertake any supportive activities with respect to the group building. However, they carefully screen the newly formed groups by imposing some restrictions: Loan contracts are denied to i) groups consisting predominantly of entrepreneurs whose income is too low to cover missing instalments of delinquent peers; and ii) groups with more than one high-risk member (entrepreneurs with bad reputation or people without permanent residence in the region of borrowing). Moreover, the loan size is decreased for applicants who had low weekly turnovers. Members of the same family household are not allowed to enter the same group.

The important difference is the following: At Constanta, the screening process takes around two weeks. During this period respective group members meet a loan officer and attend five training meetings. At FORA, the members of the newly formed groups do not go through any training sessions. The loans are disbursed within three to five days which gives the institution a high competitive advantage and considerably increases the value of the loans. At the same time, however, the simplified screening procedure reduces the amount of communication between future group members. The consequences of this difference are discussed in the next section.

In both MFIs the loan officers use the period before the actual loan disbursement to visit the applicants' businesses and evaluate

their financial situation. After the loan disbursement no official meetings with borrowers are scheduled. Nevertheless, loan officers try to stay in contact with their clients by paying them random and, as a rule, unexpected visits. If arrears occur, loan officers pay a visit to all members of the group and exert high pressure on them until the complete instalment is collected. Payments are considered delayed if one currency unit is not repaid on time. In this case, the delinquent member has to pay a fine for the late repayment and the increases in loan sizes in the next cycle are reduced for all group members.

In the history of both MFIs there were few cases of groups completely refusing to pay for a delinquent peer. These groups were immediately excluded from the lending programme and some of them were sued. The experience was that if even one defaulting group in a certain region is sentenced, the information about the process spreads immediately among all clients. Then, groups with arrears repay most of their loans very promptly.

Direct Impact of Competition: When comparing FORA with Constanta, it is obvious that the loan product of FORA has been developed in a more competitive environment, with the expected results. The product is designed in a more flexible way with respect to the cost of the product (interest rates vary with respect to several variables), the arranged repayment schedules, incentives for the clients to acquire further borrowers, incentives for the clients to stay for more than one period with the MFI, and in particular, with respect to the elapsed time between the first loan proposal by the borrower and the final approval of the loan through FORA. As mentioned in the introduction, as a consequence of the higher competitive environment of FORA its loan product appears to be of higher quality and—at least with increasing loan cycles—also less expensive.

3.4 DETERMINANTS OF THE GROUP DYNAMICS IN JOINT-LIABILITY CONTRACTS

This section addresses our research question of the impact of competition on the repayment behaviour of groups of borrowers by investigating and comparing the group dynamics in the two studied MFIs. We analyze data obtained through a survey of 241 randomly

selected credit groups of FORA and Constanta. The questionnaire studies the borrowers' socio-economic characteristics, the process of group formation, the group structure, and the intensity of the intra-group activities. All survey questions were close-ended, enabling the borrowers to give precise and unambiguous responses. To exclude correlated answers, only one randomly selected member of each group was questioned. Since the interviews were personally conducted, no serious problems with missing values occurred.

With respect to Constanta, 108 borrowers from the City of Batumi were interviewed. With respect to FORA, 133 borrowers were surveyed from four randomly selected cities: Lipetsk, Borisogletsk, Saratov, and Novgorod. The methodology used to test the hypotheses is similar to Kritikos and Vigenina.[18] To study the group-lending mechanism we constructed a multi-stage econometric model in a way that it replicates the three-stage nature of the principal-agent game: the stage of screening the borrowers before signing the loan contract, the stage of internal monitoring within the group of borrowers, and the stage of enforcing repayments. The model is separately applied to the data collected from 108 groups of Constanta and 133 groups of FORA.

3.4.1 Assortative Matching

Most theoretical research of micro-lending approaches assigns a central role to the hypothesis of assortative matching. According to this theory, borrowers with low repayment risks will group only with borrowers with similarly low repayment risks. However borrowers with high repayment risks might not be able to build stable borrower groups, even if they found other high-risk borrowers to match with.

As shown in Section 3.2, there are diverging empirical results with respect to this variable. In some MFIs borrowers were able to build homogeneous groups (and some were not). However, these approaches provide no clues to why such differing behaviour appears. In the present analysis we will therefore compare the matching behaviour of the clients of FORA where, due to the competitive environment, borrowing groups are formed in less than five days and do not participate in any training, with the behaviour of the clients

[18] Kritikos and Vigenina (2005).

of Constanta where the matching procedures take up to two weeks and where it is possible due to the low competitive pressure to ask borrowers to participate in several training sessions. By doing so, the present data set allows us to see if competition may have an impact on the process of group formation.

We start testing to see if the borrowers tend to select groups that are homogenous with respect to the variable repayment risk. To do so the interviewed borrowers were asked (in the questionnaire) how they would evaluate the aggregated repayment risk of all members of their group (excluding their own risk characteristics) at the time the group was formed (labelled *Group quality* [*GQ*]). This variable is used in the subsequent analysis as a dependent variable (see equation 1).

In a second step, in order to evaluate the risk profile of the interviewed borrower (labelled *Borrower's Risk Type* [*RT*]), we computed *RT* as a cluster analysis score (derived by Wald´s method), where the following three indicators were used: the borrower's average monthly business income, the dynamics of his income flows, and his own assessment of the stability of his business project. The surveyed clients of both MFIs were classified into three clusters with a mean value for each indicator as shown in Table 3.1. The clusters are most clearly separated by the borrowers' income level, which determines the ability of the entrepreneur to cover his instalments, thus lowering the financial risk for the lending institution. Therefore we label the

TABLE 3.1: Cluster Indicators

	FORA			*Constanta*		
	Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 1*	*Cluster 2*	*Cluster 3*
Average monthly income	5.00 (0.000)	3.93 (0.107)	1.94 (0.099)	4.64 (0.49)	2.71 (0.554)	1.5 (0.509)
Changes in the monthly income since the first loan disbursement	4.39 (0.104)	4.05 (0.13)	4.15 (0.068)	4.08 (0.49)	4.45 (0.504)	3.97 (0.186)
Stability of the business project	4.17 (0.081)	3.54 (0.079)	3.34 (0.076)	3.64 (0.7)	3.88 (0.593)	3.43 (0.504)

Source: Author.

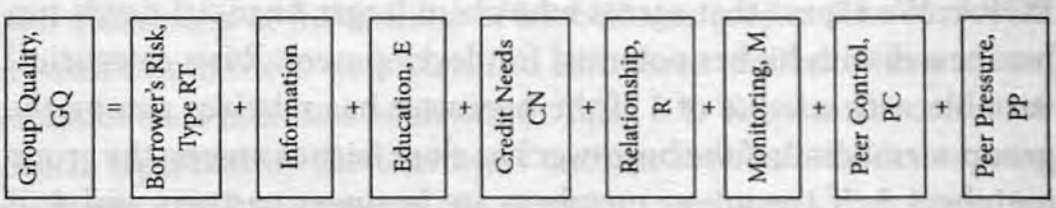
borrowers belonging to cluster 1 as 'low risk,' those from cluster 2 as 'intermediate risk,' and the rest (cluster 3) as 'higher risk.'

According to the theoretical literature the quality of self-selection depends also on how well the borrowers know their perspective peers and the business projects they run. The assumption is tested (in equation 3.1) by the second independent variable, *information (I)*—a scale ranging from 1 (if borrowers had no information about their future partners) to 5 (if they had detailed information). Tests of significance on the variables *education (E)*, *credit needs (CN)*, and *relationship (R)* are used to check whether certain personal characteristics help the applicants enter certain groups. *E* shows the borrower's level of education. It can take three possible values: 1—school, 2—college, and 3—university. *CN* indicates the amount of money the interviewed client would have borrowed, assuming that there were no constraints on the loan size. It is an interval ranging from 1 (up to 250 USD) to 5 (more than 1,000 USD) for Constanta and from 1 (up to 650 USD) to 5 (more than 2,500 USD) for FORA. We expect that agents who claim larger financial needs run businesses with higher potential for development. *R* is a categorical variable with a value of 1 if the borrower has relatives among the group members 2; if the borrower has close friends among the group members 3; if the group members are business partners and 4; if the group members are just co-signers. It is used to check whether kinship or friendship give the applicants any advantages in the selection process.

The last two variables, *peer monitoring (PM)* and *peer pressure (PP)*, test whether the internal rules in the low-risk groups substantially differ from the rules followed by the borrowers of the higher-risk groups. *PP* measures the group members' willingness to sanction delinquent partners. It is a latent variable extracted by means of the factor analysis using the following variables: 1) pressure the group (would) exert(s) on a delinquent member (answers rating from 1—no pressure—to 5—extremely strong pressure), 2) sanctions the group (would) impose(s) on a delinquent member (from 1—no sanctions—to 5—immediate exclusion from the group), and 3) sanctions the MFI (would) impose(s) on a defaulting group (from 1—the group

receives further loans but their size does not increase with time—to 5—all group members are immediately excluded from the lending program). The phenomenon of peer monitoring is presented by three variables indicating: 1) how often peers meet each other (1—once a month—to 5—every day), 2) how often they discuss their business problems within the group (1—never—to 5—on a regular basis), and 3) how well they know each other's business outcome (1—no information—to 5—very detailed information). In FORA's dataset the variables are highly correlated and load on a single latent factor, *peer monitoring*. In the case of Constanta, however, the first two variables proved to be uncorrelated and therefore enter equation (3.1) as two independent variables, called monitoring (M) and peer control (PC).

Information about the borrowers' matching behaviour can be obtained by analysing the results of equation 3.1. An ordered logit model was specified to obtain the coefficient estimates. Group quality (GQ), was defined as dependent variable.



$$\ln \frac{\sum_{k=1}^{j} P(GQ=j)}{1-\sum_{k=1}^{J} P(GQ=J)} = \alpha^{j} + \beta_1 RT + \beta_2 I + \beta_3 E + \beta_3 CN + \beta_4 R + \beta_5 M + \beta_6 PC + \beta_7 PP \quad (3.1)$$

where $j = 1, 2, \ldots, J$ indicates the *ordered* categories in the dependent variable

The significance of the explanatory variables is estimated by the likelihood ratio test. It is a joint test on all coefficients associated with each independent variable. It is assumed that a variable has no effect on Y only if all parameters associated with it are simultaneously 0. The model is calculated with and without the variable being tested. The direction of the particular variable's impact is determined by examining the results of the separate logistic regressions where the borrowers' clusters are compared in pairs (see Table 3.3) where the results from the applied multinomial logit models are also summarized.

TABLE 3.2: Multinomial Logit Model of Borrowers' Probability of Success

Likelihood Ratio Tests: Dependent variable: Borrowers' Risk Type: Cluster Score	FORA				Constanta			
Explanatory Variables	*-2Log Likelihood of Reduced Model*	*Chi-square*	*df*	*Sig*	*-2Log Likelihood of Reduced Model*	*Chi-square*	*df*	*Sig*
Intercept	129.791	0.000	0	–	104.600	0.000	0	–
Credit Needs	131.170	1.379	2	0.502	105.635	1.035	2	0.596
Self-selection	132.919	3.128	2	0.209	114.864	10.264	2	0.006
Education	134.113	4.321	4	0.364	109.080	4.480	4	0.345
Nr. of borrowers in cluster 1				36				22
Nr. of borrowers in cluster 2				52				40
Nr. of borrowers in cluster 3				21				29
Pseudo R2 =				.083				.178
Model significance (test the joint significance of all independent variables):								
Chi-Square value =				8.172				15.512
df =				8				8
P-value =				.417				.050

Source: Author.

Notes: This table presents the results from the likelihood ratio tests of hypothesis 1. The likelihood ratio test is a joint test on all coefficients associated with each exogenous variable. The model is calculated with and without the variable being tested. The first column of the table lists the exogenous variables. The second column contains the—2LL values of the reduced model. The latter is formed by omitting a variable from the final model. The null hypothesis is that all parameters of that variable are 0. The third column presents the chi square statistic: the difference in –2 log-likelihoods between the final model and a reduced model. Degree of freedom (df) is the difference in the number of parameters in the final model and the reduced model.

TABLE 3.3: Multinomial Logit Model of Borrowers' Matching Behaviour (separate logistic regressions with pairwise comparisons)

Outcome Variables:	*Cluster 1 vs. Cluster 3*	*Cluster 2 vs. Cluster 3*	*Cluster 1 vs. Cluster 2*
Explanatory Variables:	*Coefficient (Wald Statistic)*	*Coefficient (Wald Statistic)*	*Coefficient (Wald Statistic)*
Intercept	-9.990***	-4.937***	-5.054*
	(9.051)	(3.133)	(3.534)
Credit Needs	0.000	0.000	-0.000
	(0.456)	(0.023)	(0.911)
Self-Selection	2.195***	1.271***	0.924
	(8.224)	(3.591)	(2.466)
Education (1)	21.427***	19.684	1.743
	(236.739)	(.)	(1.566)
Education (2)	0.403	0.124	0.279
	(0.260)	(0.041)	(0.149)
Education (3)	0.000[a]	0.000[a]	0.000[a]
	(.)	(.)	(.)

Source: Author.

Notes: *** Significant at the 1 per cent level; ** Significant at the 5 per cent level; * Significant at the 10 per cent level.

[a] This parameter is set to zero because it is redundant.

The table presents the parameter estimates from the separate logistic regressions of the multinomial logit model used for testing equation 1. We compare, in succession, 1) the low risk borrowers (cluster 1) with higher risk borrowers (cluster 3), 2) intermediate risk borrowers (cluster 2) with higher risk borrowers (cluster 3), and 3) low risk borrowers (cluster 1) with intermediate risk borrowers (cluster 2). The statistical significance of the individual coefficients is evaluated by the Wald test. The Wald statistic for the test of a single coefficient is contained in parentheses. It equals the square of the ratio of the coefficient divided by its standard error and asymptotically follows a chi-square distribution:

$$W_k^2 = \left(\frac{b_k}{s_{b_k}}\right)^2 - \chi^2.$$

For Constanta's groups the model is well specified (see Table 3.2): The test of joint significance of all independent variables has a *p*-value of .05. *Self-selection* is the only significant variable (at 99 per cent confidence level), confirming the assortative matching through group lending. The insignificance of all other independent factors shows that these do not contribute to the separation of the clusters and thus cannot be used by the loan officers to differentiate between borrowers' risk types.

In FORA's groups, however, we could not find evidence of homogeneous matching (see Table 3.2). The high p-value (.417) of the provided test of the overall model indicates the bad explanatory power of the whole set of independent factors. The heterogeneity of the groups could be partly explained by the short self-selection process, as a result of which only a relatively small amount of initial information about the respective peers is accumulated.

Whether the assumption of assortative matching holds depends not only on the endogenous process of group formation but also on the screening policy of the MFI—whether it makes respective borrowers attend training sessions during the selection process or uses short straightforward screening procedures, thus stressing on the speedy disbursement of loans. Apart from the freedom to choose their partners independently, the borrowers need sufficient time and even encouragement by the loan officers to gather the necessary information about their future partners, and subsequently to self-select in homogeneous in respect to the investment risk groups.

Result 1: At the MFI Constanta the group formation is influenced by the variable *Borrower's risk type*. We found that lower risk borrowers indeed team up with lower risk borrowers and vice versa. For the borrowing groups of FORA we found no evidence of assortative matching with respect to the risk attitudes of the borrowers.

3.4.2 Testing the Efficiency of the Applied Incentive Mechanisms

Once loans are disbursed, the lender's success depends on the repayment behaviour of both the groups and (indirectly) the individual borrowers. The second stage of the principal-agent game is presented by a system of two equations. Equation 3.2 tests the factors that may motivate the individual borrowers to promptly fulfill their repayment obligations, i.e., choose safe projects or not divert means for consumption purposes. It accounts for the repayment performance inside the group. Equation (3.3) reveals the determinants of the repayment behaviour of the entire group.

Both dependent factors are dichotomous variables taking a value of 1 if there were no cases of internal or external delinquency in the

group and 0 otherwise. To consider the effect of the independent factors binary logit models are specified.

Internal Repayment Performance, Y_2	=	Peer Monitoring, PM	+	Peer Pressure, PP	+	Peer Support, PS	+	Dyn. Incentives, DI	+	Borrower's Risk Type Y_1	+	Staff Pressure, SP	+	Business Correlation, CORR	+	Social Ties, ST

$$\ln \frac{P(Y_2 = 1)}{1 - P(Y_2 = 1)} = \alpha + \gamma_1 PM + \gamma_2 PP + \gamma_3 PS + \gamma_4 DI + \gamma_5 Y_1 + \gamma_6 SP + \gamma_7 CORR + \gamma_8 ST \quad (3.2)$$

where $Y_2 = 1$ if there was no incidence of internal delinquency in the credit group.

External Repayment Performance, Y_3	=	Peer Monitoring, PM	+	Peer Pressure, PP	+	Peer Support, PS	+	Dyn. Incentives, DI	+	Self-Selection S	+	Staff Pressure, SP	+	Business Correlation, CORR	+	Social Ties, ST

$$\ln \frac{P(Y_3 = 1)}{1 - P(Y_3 = 1)} = \alpha + \delta_1 PM + \delta_2 PP + \delta_3 PS + \delta_4 DI + \delta_5 S + \delta_6 SP + \delta_7 CORR + \delta_8 ST \quad (3.3)$$

where $Y_3 = 1$ if there was no incidence of external delinquency in the credit group.

The use of two equations allows not only to address the complexity of the group-lending repayment system better, but also helps to understand the incentive mechanism process: e.g., if they prevent the occurrence of repayment failures, thus mitigating the moral hazard problem, or serve as corrective measures, tackling the enforcement problem only.

The explanatory factors in both equations are identical. Some of them reflect the special features of the lending mechanism; others account for the lender's characteristics or reflect specific socioeconomic conditions. To the first group belong peer support, peer monitoring, peer pressure, and dynamic incentives. The borrower's risk type (in equation [3.2]) and the self-selection (in equation [3.3]) are used as a measure of the quality of the clients' portfolio: The repayment performance of the borrowers is influenced by their individual characteristics, whereas the repayment performance of the complete group depends on characteristics that are common for all group members.

Business correlation reflects the degree of positive correlation across borrowers' projects' returns. Social ties measures the homogeneity of the groups in terms of age, gender, income, etc. Staff pressure indicates the impact of the loan officers' monitoring efforts. The last factor, loan term to maturity, is thought to be an effective monitoring tool that helps the MFIs discipline the clients and gives the loan officers early warning about emerging problems.

To estimate the coefficients of the independent variables, we apply the method of maximum likelihood. The parameter significance is evaluated by the Wald statistics, which equals the square of the ratio of the coefficient divided by its standard error and asymptomatically follows a chi-square distribution:

$$W_k^2 = \left(\frac{b_k}{s_{b_k}}\right)^2 \sim \chi^2.$$

The interpretation is similar to the t values used for the significance testing of linear regression coefficients.

The instruments for managing the credit risk used by the studied MFIs are, with some small exceptions, identical. The impact of these instruments on the borrowers' behaviour, however, proved to differ most probably due to the revealed differences in the group formation and structure. In the following paragraph, we highlight the role of the incentive mechanisms created by the joint-liability contract in solving the delinquency problems. Our aim is to show what motivates the clients to both invest the borrowed capital into safe projects and avoid strategic defaults.

We start with analysing the factors that lead to the improvement of the repayment performance of individual borrowers. The empirical results from the specified binary logit model (equation (3.2)) are presented in Table 3.4. The only variable that proved to be positive and statistically significant in both samples of borrowers is *peer support*. It measures the group members' willingness for mutual support. The belief of the borrowers that they will be helped in case of an external shock makes them perceive the group as a reliable source of insurance, thus implicitly increasing the value of the group loan.

Further we investigate how the strength of the peer support changes with time when the differences among the group members are

TABLE 3.4: Binary Logit Model of Group Internal Repayment Performance

	Dependent Variable: Internal Repayment Performance					
	FORA			*Constanta*		
Explanatory Variables:	*Coefficient B (Standard Error)*	*Wald Statistics*	*Exp (B)*	*Coefficient B (Standard Error)*	*Wald Statistics*	*Exp (B)*
Peer Support	0.508* (0.285)	3.165	1.661	2,649** (1,159)	5,227	14,14
Dynamic Incentives	0.493 (0.415)	1.408	1.637	−0,091 (0,369)	0,061	0,913
Borrower's Risk Type		1.089			1,773	
Borrower's Risk Type	0.968 (0.928)	1.087	2.632	−0,914 (0,844)	1,173	0,401
Borrower's Risk Type	0.687 (0.888)	0.599	1.988	0,025 (0,76)	0,001	1,025
Peer Pressure	−0.366 (0.383)	0.917	0.693	0,472 (0,393)	1,446	1,603
Monitoring	0.743** (0.366)	4.135	2.103	0,068 (0,355)	0,037	1,071
Controlling				−0,748*** (0,246)	9,236	0,473
Staff Pressure	0.075 (0.296)	0.063	1.078	−0,399 (0,266)	2,257	0,671
Business Correlation	−0.152 (0.272)	0.313	0.859	−0,007 (0,27)	0,001	0,993
Social Ties	−0.574*** (0.206)	7.768	0.563	−0,03 (0,286)	0,011	0,971
Loan Term	−0.715** (0.297)	5.795	0.489			
Intercept	4.076* (2.078)	3.847	58.934	4,071* (2,203)	3,416	58,596
Groups with internal repayment problems = 0			24			21
Groups without internal repayment problems = 1			53			66
Pseudo R2 =			0.340			0.374
Model significance (test the joint significance of all independent variables):						
Chi-Square =			21.294			25.057
df =			10			10
P-value =			0.019			0.005

Source: Author.

Notes: ***Significance at the 1 per cent level; **Significance at the per cent level
The table presents the results from a binary logit model of group internal repayment performance (equation 2). The coefficients (column 2) are maximum likelihood estimates. The statistical significance of the individual coefficients is evaluated by the Wald test. The Wald statistic for a test of a single coefficient (column 3) equals the square of the ratio of the coefficient divided by its standard error and asymptotically follows a chi-square distribution:

$$W_k^2 = \left(\frac{b_k}{s_{b_k}}\right)^2 \sim \chi^2.$$

Column 4 lists the odd ratios (Exp B).

expected to grow bigger, causing serious mismatching problems. The set of independent variables is regressed on the internal repayment performance of both the 'new' borrowers (with three or less loans) and the 'old' borrowers (with more than three loans).

In Constanta, peer support is significant only in the sub-sample of 'old' borrowers, evidencing that group support grows stronger in the course of time. Contrary to prior expectations, the cooperation among borrowers improves the longer they stay in the group. It shows that using the group-lending technology, Constanta will be very likely to generate high repayments over a relatively long period of time. FORA's empirical results show the reverse tendency. In this case peer support is significant only in the sub-sample consisting of new borrowers, evidencing that the belief in the group willingness to provide mutual help improves the repayment rate only during the first loan cycles and significantly diminishes afterwards.

Result 2: In both MFIs the variable peer support turned out to be a crucial intra-group variable helping to solve repayment problems. However, at FORA, peer support is effective during the first three loan cycles when borrowers barely know each other. At Constanta, peer support is only provided in groups who were together for more than three loan cycles. The result indicates that borrowers of FORA offered peer support until they stopped trusting each other, while borrowers of Constanta started to trust each other after they created a group relationship.

The repayment rate of FORA could be expected to decrease constantly with time. According to the official financial analysis, however, it has significantly improved over the last five years.

Apparently, there are other incentive mechanisms that help the MFI mitigate the moral hazard problem. One such mechanism proved to be peer monitoring. It is computed as a factor analysis score indicating 1) how often peers meet each other, 2) how often they discuss their business problems within the group, and 3) how well they know one another's business outcome. The variable is highly significant and positive, evidencing that the more intensively the peers monitor each other the less likely the internal delinquency is. FORA's clients effectively use the meetings to control their partners and reassure that the latter do not misallocate the borrowed capital.

In case of Constanta the phenomenon of peer monitoring is presented by two variables called monitoring and controlling (see also Table 3.4). Monitoring shows the frequency of meeting between the peers and is thus a proxy for the borrower's ex ante peer-monitoring efforts. Controlling indicates how often the borrowers discuss their business problems within the group. Our initial expectations were that the two variables would load on a single factor as in the case of FORA but they proved to be statistically unrelated and were included into the model as two independent variables. Controlling is highly significant (at 99 per cent confidence level) but surprisingly has a negative coefficient. This can be explained by the fact that in the more problematic groups the intra-group communication is intensified, showing that peers try to verify the real state of the projects only if some of them fail to repay.

The next factor to be discussed is social ties (an index measuring the homogeneity of the groups with respect to age, gender, income, etc.). It is believed that the strength of the social ties between the group members strongly affects their behaviour as well as their exposure to various problems. According to the theory, the repayment performance of a group should be better when the social ties between the members are strong since the latter increase the mutual trust and at the same time strengthens the impact of the imposed social sanctions.

Our results contradict this assumption by showing that social ties have no (at Constanta) or rather a negative (at FORA) effect on borrowers' repayment behaviour. At FORA, the variable negatively influences the willingness (or ability) of both individual borrowers and groups to repay the loan: It is statistically significant with a negative

coefficient in equation (3.2) as well as equation (3.3). To explain this phenomenon we assert that in socially homogenous groups the clients rely greatly on their peers' unconditional support and are thus more likely to free ride.

In FORA's groups, the willingness for mutual financial support (peer support significantly improves both internal and external repayment performance) substantially increases the chances of a delinquent borrower to be refinanced by the lender despite his repayment failure. However, as the empirical results show, borrowers not only explicitly express their intention to support their peers but also effectively monitor them (peer monitoring is statistically significant at 95 per cent confidence level), sufficiently raising the probability of discovering the real reason for the default. Thus, free riding, being very likely to be discovered, often creates tension and mistrust among the group members, resulting in a group refusal to cover the debt of the fraud (explaining the negative sign of social ties in Table 3.4).

Except for joint liability, the majority of the group-lending MFIs also use other mechanisms that are expected to discipline the borrowers and improve the repayment performance. One of them is the shorter term to maturity. It is thought to be an effective monitoring tool, especially in the first loan cycles. Longer loans are considered to be riskier since they have more chances to fall into arrears and lead to greater delinquency rates.

To examine the role of the loan duration, we slightly extend the econometric model by adding a new variable called loan term. The modified model is applied only to FORA's database since Constanta offers only one type of loan contract to all of its clients, where the time to maturity is fixed at exactly four months. FORA's lending procedure is more flexible in the sense that it allows the loan duration to vary from client to client in accordance to the cash flows. It ranges from four to twelve months. Table 3.3 shows that loan term is statistically significant (at 95 per cent confidence level) and, as expected, displays a negative sign. The longer the term to maturity, the more difficult it is for the borrowers to meet the repayment obligations.

Result 3: (1) For the MFI FORA, we found that group members effectively monitor each other to keep the probability of repayment failure at a low level. In contrast to this, borrowers of Constanta start

monitoring each only when a repayment problem occurred. (2) Social ties have no (at Constanta) or even a negative (at FORA) effect on borrowers' repayment behaviour. (3) The flexible loan terms developed at FORA have a negative impact on repayment behaviour.

Despite all precautionary measures, some borrowers fail to fulfill their repayment obligations on time. Repayment problems occur as a result of both external uncontrollable shocks and clients' carelessness. Notwithstanding the reason, the credit groups are expected to always repair the individual's failure either by covering the delinquent person's part of the loan or by imposing (social) sanctions, thus forcing them to repay. In this sub-section we attempt to reveal the variables mostly influencing the repayment performance of the credit groups of the two MFIs. Whether the groups repair the repayment failures of individual borrowers, and what kind of peer measures they use is investigated by regressing the external repayment performance on the previously specified set of independent variables.

TABLE 3.5: Binary Logit Model of Group External Repayment Performance

	Dependent Variable: External Repayment Performance					
	FORA			*Constanta*		
Explanatory Variables:	*Coefficient B (Standard Error)*	*Wald Statistics*	*Exp(B)*	*Coefficient B (Standard Error)*	*Wald Statistics*	*Exp(B)*
Peer Support	0.620* (0.336)	3.414	1.860	16,023 (58,11)	0,076	9096945
Dynamic Incentives	0.646 (0.489)	1.746	1.908	−0,826 (0,985)	0,702	0,438
Self-Selection	0.601 (0.708)	0.721	1.825	8,583** (3,983)	4,643	5341,925
Peer Pressure	−0.972** (0.471)	4.264	0.378	3,585** (1,749)	4,204	36,065
Monitoring	0.338 (0.420)	0.649	1.403	−0,231 (0,884)	0,068	0,794
Controlling				−2,611** (1,214)	4,623	0,073
Staff Pressure	0.040 (0.359)	0.013	1.041	1,032 (0,664)	2,411	2,806

(Contd.)

(Table 3.5 contd.)

	Dependent Variable: External Repayment Performance					
	FORA			*Constanta*		
Explanatory Variables:	*Coefficient B (Standard Error)*	*Wald Statistics*	*Exp(B)*	*Coefficient B (Standard Error)*	*Wald Statistics*	*Exp(B)*
Business Correlation	−0.526* (0.319)	2.720	0.591	−2,263* (1,295)	3,054	0,104
Social Ties	−0.717*** (0.237)	9.195	0.488	−1,227 (0,863)	2,023	0,293
Loan Term	−1.051*** (0.364)	8.329	0.350			
Intercept	5.396 (3.525)	2.343	220.438	−13,13 (12,183)	1,162	0
Groups with internal repayment problems = 0			22			8
Groups without internal repayment problems = 1			49			78
Pseudo R2 =			0.463			0.770
Model significance (test the joint significance of all independent variables):						
Chi-Square =			28.302			37.786
df =			9			9
P-value =			0.001			0.000

Source: Author.

Notes: ***Significance at the 1 per cent level; **Significance at the 5 per cent level; *Significance at the 10 per cent level

The table presents the results from a binary logit model of group external repayment performance (equation 3). The coefficients (column 2) are maximum likelihood estimates. The statistical significance of the individual coefficients is evaluated by the Wald test. The Wald statistic for a test of a single coefficient (column 3) equals the square of the ratio of the coefficient divided by its standard error and asymptotically follows a chi-square distribution:

$$W_k^2 = \left(\frac{b_k}{s_{b_k}}\right)^2 \sim \chi^2.$$

Column 4 lists the odd ratios (Exp B).

Table 3.5 reveals substantial differences between FORA and Constanta in terms of the intra-group behaviour of their clients. One crucial variable seems to be peer pressure which is statistically

significant in both samples of borrowers but the coefficient takes different signs. In Constanta, peer pressure proved to positively influence the repayment decisions of group members, indicating that the probability of default for groups whose members impose (or express their intention to impose) social sanctions is significantly smaller. The fact that the variable improves only the external repayment rate and does not affect the internal repayment performance allows for the interpretation that the majority of Constanta's borrowers exert ex post pressure. They start acting after an internal problem has occurred. The negative sign of the controlling variable (significant at 95 per cent confidence level) can be seen as a further confirmation of this explanation.

The behavioral pattern of FORA's borrowers is much different. In FORA, peer pressure is also significant but has a negative coefficient. We assert that only clients who belong to groups who have already been at least once in arrears are willing to sanction a delinquent member. These are mainly groups where the tension between the peers is high as they have not successfully managed to solve their internal problems. The clients of FORA, different from those of Constanta, rarely use peer pressure for inducing repayments. They apparently employ it at a later stage—after they have paid for the delinquent partner—in order to get compensated for the payments they made for their delinquent peers in the group.

Instead, the most efficient incentive mechanism for tackling the enforcement problems in FORA turned out to be peer support. It is the only statistically significant and positive variable in equation (3.3). The group serves as a secondary repayment source for the clients and thus protects the lender from suffering losses due to repayment failures.

Another factor with different implications on the borrowers of the two MFIs is self-selection (see also Table 3.5). It is of no surprise that the factor does not influence the repayment performance of FORA's groups, taking into account the fact that no evidence of homogeneous matching has been found. In contrast to this, for Constanta, the quality of the selection process is important for its successful financial development. We have shown that the individual characteristics, measured by the borrower's risk type, are insignificant for the

improvement of the internal repayment performance. It means that, for instance, middling risk borrowers do not default more often within their groups than low risk clients. The significance of self-selection, however, indicates that groups consisting of higher risk borrowers are more likely to be delinquent than groups formed by clients with low risk characteristics.

The last point to be discussed is the negative impact of the correlation across the group members' businesses. Constanta and FORA serve predominantly traders. As a result most of the groups are homogeneous with respect to economic activities or risk exposure. The degree of such homogeneity is measured by the variable business correlation. The borrowers were asked if they were involved in mutual trading activities, had common clientele, or common supplier. The priori expectations about this variable are ambiguous. On the one hand, high business correlation is associated with bad risk diversification in the group and increases the probability of failure. On the other hand, however, the higher degree of positive correlation increases the borrowers' incentives to monitor each other and thus raises the probability of success of all projects.[19] In order to check the validity of the latter statement we regress business correlation on monitoring. The coefficients are highly insignificant (Constanta: p-value = 0.870; FORA: p-value = 0.411) showing that peers with correlated businesses do not excessively intensify their monitoring efforts in order to avoid paying the debts of their partners. However, in both MFIs, business correlation proved to be significant with a negative coefficient. Business dependence across group members substantially weakens the peers' ability to mutually insure each other when external shocks arise, thus increasing the likelihood of a group default.

Result 4: (1) Among the borrowers of Constanta, the variable peer pressure significantly improves the external repayment, indicating that borrowers exert peer pressure if a repayment problem has occurred. Moreover, the self-selection process has a positive impact on repayment when low-risk borrowers grouped together. (2) Among the borrowers of FORA, peer pressure is only used by borrowers who paid for delinquent co-signers and who aim to get compensated from these

[19] Armendariz de Aghion (1999).

delinquent peers. (3) Business correlation within a group of borrowers increases the probability of repayment failure in both MFIs.

3.5 IMPACT OF COMPETITION BETWEEN MFIs ON LOAN GROUPS

This paper studies the group dynamics of two MFIs which provide access to micro-loans by making use of joint-liability contracts. As competition between MFIs in India has become an important topic, the main aim of the paper is to find out whether competition between suppliers of micro-loans has an impact on these group dynamics. As there was no sufficient access to data in India allowing to test the main hypothesis, we decided to analyse the research question by making use of data from MFIs situated in republics of the former Soviet Union. We compared the observed behaviour in loan groups of Constanta (an MFI in Georgia) which had no serious competitor in the field, with the behaviour of borrowers of FORA (an MFI in Russia) which faced strong client competition. Due to the competitive environment, the applicants of FORA had also access to loans with lower interest rates (in comparison to Constanta) and with more flexibility, such as a longer term of maturity. As most other variables as well as the cultural setting in the two MFIs are similar, we were able to make a comparison.

Our results indicate that competition has a significant impact on the intra-group behaviour of borrowers. Although we revealed that the group-lending mechanism proved to be effective in inducing on-time repayments in both institutions, we detected substantial differences in the way it works. We had structured the complete joint-liability contract into three different stages, the first stage being the group matching, the second being the behaviour within the group after contracts were signed, and the third stage being the time of due loan repayment.

Starting with the process of group formation we found that giving the applicants the ability to independently self- and co-select does not automatically lead to assortative matching. A careful matching occurs only if the clients have the option (and probably also are encouraged by the loan officers) to gather detailed information about their peers. According to the empirical results, such an approach—applied in Constanta—helps the lender to alleviate the adverse selection problem, but requires a selection period of about two weeks. In a competitive

environment the slow screening procedure is known to impede the growth of the credit institutions as it makes them less attractive for the entrepreneurs. Accordingly, the matching procedure in FORA took place within three to five days. As a consequence, we found out that the applicants of Constanta had matched together to groups of homogenous risk types, while the applicants of FORA had not.

In the next steps, we were able to observe the consequences of the differing group matchings. The analysis shows that in the case of FORA, borrowers compensate for the lack of sufficient initial intra-group information by carefully monitoring their peers during the entire lending period. They not only regularly meet their peers but also control their business outcomes. Constanta's borrowers, on the other hand, know their future peers already, rely on their screening efforts and do not monitor one anothers' business activities as long as there are no signals of repayment failures. When a problem occurs they intensify their peer monitoring by auditing the defaulter's work.

This means that the cost of information which the borrowers of Constanta had already 'invested' during the screening process, were not 'saved' by the borrowers of FORA. They simply had to do this investment later on and we cannot exclude that the monitoring efforts had been higher among the borrowers of FORA when compared to Constanta because instead of a one time screening process (over two weeks) a constant monitoring might have been necessary as borrowers with different risk levels were grouped together.

Besides peer monitoring, we also found out that another important precondition for the success of group lending in both MFIs is the ex ante willingness for peer support. It signals the borrowers' readiness for intra-group cooperation. The strong belief in the mutual support considerably increases the value of the group loan and leads to mitigation of the moral hazard problem. However, as we showed in the comparison of Constanta with FORA, there exist different ways of peer support conditional on the quality of the selection process. The biggest difference between the MFIs is that in Constanta the feeling of group solidarity grows stronger in the course of time. In these groups persons can be relatively sure that the payments for the next instalments of their peers will be returned by the delinquent borrowers as soon as possible. In FORA group solidarity starts diminishing after

the first three loan cycles. A possible reason for the growing mistrust among FORA's clients could be the fact that even though the most group members stay in close relationships, they typically lack prior experience in financial matters with one another. Only about 4 per cent of the interviewed borrowers describe their peers as business associates, while 44 per cent define them just as cosigners and over 50 per cent as friends. As the loan officers report, many borrowers falsely believe that people who proved to be good friends will be good business partners as well. When the expectations are not met, mistrust destroys the work of the group as their delinquent peers for whom they provided financial support obviously refused to repay. To compare, in Constanta 47 per cent of the groups are formed by business associates, around 21 per cent by distant relatives, 30 per cent by friends, and only 2 per cent by cosigners. Thus, the missing screening efforts have a costly impact on peer support as well.

In Constanta, peer pressure also proved to be an additional effective instrument to enforce repayments. Similar to peer monitoring, peer pressure is used by the borrowers only as a corrective measure. Worth noting is the fact that in Constanta the willingness of the group to pay for delinquent members is highly correlated with the willingness to exert pressure on them. We assume that the motivation of Constanta's clients to pay for their partners is not simply the willingness to support them but rather the willingness to promptly fulfil the repayment obligations. In contrast to this, among borrowers of FORA, peer pressure took place within the groups only to get compensated for the payments made for delinquent peers. Accordingly, peer pressure was also not used to enforce repayments within the group when the next instalment was due.

Last but not least, social cohesion and business correlation are the only two variables whose impact on the borrowers' behaviour does not vary between the two MFIs. According to the empirical analysis, the probability of default in highly cohesive groups is significantly higher than in groups with no social cohesion. Apparently, in such groups borrowers stronger rely on the peers' unconditional support and are thus more likely to free ride. Moreover, in both MFIs it appeared that the probability of default within a group was significantly higher when the businesses were correlated. This observation shows that it is an

important task for the loan officers to accept only groups of borrowers where the businesses are sufficiently diversified.

Putting all observations of the group dynamics in the studied MFIs together, we conclude that the self-selection process is a crucial activity between borrowers. The quality of self-selection determines the intensity of all further peer measures. Moreover, the empirical results show that the main factors driving the high repayment rate are the peer activities induced by the principle of joint liability. In this context, the finding that the group-lending mechanism induces prompt loan repayments even if borrowers group randomly (FORA), is of importance as well. It makes clear that for a certain—rather short—time period groups may work in the intended way of joint-liability even if the members of the groups were not able to realize an assortative matching.

However, it comes at significant cost that joint liability works even if groups matched randomly because of a competitive environment. It is necessary to highlight in particular three negative impacts of competition on joint-liability contracts. The first disadvantage is the lower stability of randomly matched groups. Most of them do not survive more than three loan cycles (while in the MFI Constanta groups started after three loan cycles to work together in a better way). Because of the negative experiences with delinquent peers, members of these groups have a decreasing motivation to stay on good terms with their peers. FORA found a solution which enables them neither to lose these borrowers nor to face increasing repayment problems; they offer to the clients the option to switch to an individual lending contract after a few loan cycles. The second disadvantage of joint-liability contracts under competition is that the efforts which these borrowers safe during the self-selection process of the groups (and which are invested by the borrowers of Constanta) might be over-compensated by the later efforts of additional peer monitoring and peer pressure. The third disadvantage is that borrowers who signed a joint liability contract in a competitive environment and who grouped randomly face higher cost when they decide to provide delinquent peers with financial support for which they are not necessarily compensated.

We conclude therefore that competition may reduce the direct costs (for interest payments) of loans which were given under joint liability.

However, these lower direct costs are most probably overcompensated by additional efforts of peer monitoring and peer pressure, as well as by the additional costs when group members provide peer support to delinquent members. Our results indicate therefore that joint-liability loan contracts might be less stable and more expensive in a competitive environment even though their interest rates are lower.

REFERENCES

Armendariz de Aghion, B. (1999), 'On the Design of a Credit Agreement with Peer Monitoring', *Journal of Development Economics*, vol. 60, pp. 79–104.

Banerjee, A., T. Besley and T. Guinnane (1994), 'Thy Neighbor's Keeper: The Design of a Credit Cooperative', *Quarterly Journal of Economics*, vol. 109, pp. 491–515.

Besley, T. and S. Coate (1995), 'Group Lending, Repayment Incentives and Social Collateral', *Journal of Development Economics*, vol. 46, pp. 1–8.

Ghatak, M. (1999), 'Group Lending, Local Information and Peer Selection', *Journal of Development Economics*, vol. 60, pp. 27–50.

Hoff, K. and J. Stiglitz (1990), 'Imperfect Information and Rural Credit Markets—Puzzles and Policy Perspectives', *The World Bank Economic Review*, vol. 4, pp. 235–50.

Hulme, D. and P. Mosley (1996), *Finance Against Poverty*, London: Routledge.

Khandker, S. (1998), *Fighting Poverty with Microcredit: Experience in Bangladesh*, New York: Oxford University Press for the World Bank.

Kritikos, A.S. (1999), 'Micro-credits with incentive-guided solidarity action', *Kredit und Kapital*, vol. 32, pp. 393–425.

Kritikos, A.S., Vigenina, D. (2005), 'Key Factors of Joint-Liability Loan Contracts', *Kyklos*, vol. 58, pp. 213–38.

Laffont, J.J. and T. N'Guessan (2000), 'Group Lending with Adverse Selection', *European Economic Review*, vol. 44, pp. 773–84.

Microfinance Information Exchange (MIX) (2007), 'India Microfinance Review', Washington.

McIntosh, C., A. de Janvry and E. Sadoulet (2004), 'How Rising Competition Among Microfinance Institutions Affects Incumbent Lenders', Disc. Paper, San Diego.

Morduch, J. (1999), 'The Microfincance Promise', *Journal of Economic Literature*, vol. 37, pp. 1569–1614.

Sadoulet, L. and S. Carpenter (1999), 'Risk-Matching in Credit Groups: Evidence from Guatemala', *ECARES, Free University of Brussels*.

Sharma, M. and M. Zeller (1997), 'Repayment Performance in Group-Based Credit Programs in Bangladesh: An Empirical Analysis', *World Development*, vol. 25.

Stiglitz, J.E. (1990), 'Peer Monitoring and Credit Markets', *The World Bank Economic Review*, vol. 4, pp. 351–66.

Van Tassel, E. (1999), 'Group Lending Under Asymmetric Information', *Journal of Development Economics*, vol. 60, pp. 3–25.

Varian, H.R. (1990), 'Monitoring Agents with Other Agents', *Journal of Institutional and Theoretical Economics*, vol. 146, pp. 153–74.

Vigenina, D. and A.S. Kritikos (2004), 'The Individual Micro-Lending Contract: Is it a Better Design than Joint-Liability?' *Economic Systems*, vol. 28, pp. 155–76.

Wenner, M.D. (1995), 'Group Credit: A Means to Improve Information Transfer and Loan Repayment Performance', *Journal of Development Studies*, vol. 32, pp. 263–81.

Wydick, B. (1999), 'Can Social Cohesion be Harnessed to Repair Market Failure? Evidence from Group Lending in Guatemala', *Economic Journal*, vol. 109, pp. 463–75.

Sharma, M. and M. Zeller (1997), 'Repayment Performance in Group-Based Credit Programs in Bangladesh: An Empirical Analysis', World Development, vol. 25.

Stiglitz, J.E. (1990), 'Peer Monitoring and Credit Markets', The World Bank Economic Review, vol. 4, pp. 351–66.

Van Tassel, E. (1999), 'Group Lending Under Asymmetric Information', Journal of Development Economics, vol. 60, pp. 3–25.

Varian, H.R. (1990), 'Monitoring Agents with Other Agents', Journal of Institutional and Theoretical Economics, vol. 146, pp. 153–74.

Vigenina, D. and A.S. Kritikos (2004), 'The Individual Micro-Lending Contract: Is it a Better Design than Joint-Liability? Economic Systems, vol. 28, pp. 155–76.

Wenner, M.D. (1995), 'Group Credit: A Means to Improve Information Transfer and Loan Repayment Performance', Journal of Development Studies, vol. 32, pp. 263–81.

Wydick, B. (1999), 'Can Social Cohesion be Harnessed to Repair Market Failure? Evidence from Group Lending in Guatemala', Economic Journal, vol. 109, pp. 463–75.

II

ACCIDENT LAW AND ENVIRONMENTAL PROBLEMS

4

NEGLIGENCE RULE

Some Strategic Aspects

Satish K. Jain

One of the most important liability rules, the rules for apportioning accident loss between victim and injurer, is the negligence rule. Under the negligence rule injurer is liable for the entire accident loss iff he is negligent; and he is not liable for any part of accident loss iff he is non-negligent. In tort law the notion of negligence is defined in several different but related ways.

One of the most important results of law and economics literature, and one of the earliest, relates to the efficiency of negligence rule.[1] If (1) negligence is defined as failure to take a level of care greater than or equal to a specified level of care, called due care, and (2) due care is set at a level at which social costs (sum of precaution costs and expected accident loss) are minimized; then under the negligence rule

[1] One of the earliest analyses of negligence rule from the efficiency perspective was carried out in R.A. Posner (1972), 'A Theory of Negligence', *Journal of Legal Studies*, vol. 1, pp. 28–96. The efficiency of negligence rule was formally established by J.P. Brown (1973), 'Toward an Economic Theory of Liability', *Journal of Legal Studies*, vol. 2, pp. 323–50. On the negligence rule generally see W.M. Landes and R.A. Posner (1987), *The Economic Structure of Tort Law*, Cambridge, MA: Harvard University Press, S. Shavell (1987), *Economic Analysis of Accident Law*, Cambridge, MA: Harvard University Press, R.D. Cooter and T.S. Ulen (2003), *Law and Economics*, 4th edn, New York: Addison-Wesley, and R.A. Posner (2007), *Economic Analysis of Law*, 7th edn, New York: Wolters Kluwer Law & Business, among others.

both victim and injurer are induced to take care levels which would minimize social costs.

Notion of negligence at times is defined without any reference to a specified due care level. Injurer is adjudged to be negligent iff there exists a cost-justified untaken precaution; and non-negligent iff there does not exist any cost-justified untaken precaution.[2] A precaution is called cost-justified iff taking of it increases cost of care by an amount less than the reduction in expected loss as a result of taking it. Although negligence rule is generally thought to be efficient under this notion of negligence also; it is not so.[3]

This paper is concerned with some important strategic aspects which the negligence rule gives rise to; and which have been largely neglected in the analysis pertaining to efficiency. When these strategic aspects are brought into consideration explicitly the efficiency analysis changes significantly as these strategic aspects tend to work against the objective of minimizing social costs of accidents.

This paper is divided into four sections. The first section contains the preliminaries. The second section analyses the strategic aspects of the negligence rule when the notion of negligence is defined as shortfall from due care level; and due care level is specified from the perspective of minimization of social costs. The next section is concerned with analysing the strategic aspects of the negligence rule when the notion of negligence is defined as existence of a cost-justified untaken precaution. The concluding section, in addition to containing a summary of the main conclusions, points out that the problem of strategic manipulation is not confined to the negligence rule but extends to all efficient liability rules. The section also contains an intuitive explanation for the strategic vulnerability of efficient liability rules.

[2] On the notion of negligence as existence of a cost-justified untaken precaution, see Grady (M.F. Grady (1983), 'A New Positive Theory of Negligence', *Yale Law Journal*, vol. 92, pp. 799–829; M.F. Grady (1984), 'Proximate Cause and the Law of Negligence', *Iowa Law Review*, vol. 69, pp. 363–449; M.F. Grady (1989), 'Untaken Precautions', *Journal of Legal Studies*, vol. 18, pp. 139–56).

[3] See S.K. Jain (2006), 'Efficiency of Liability Rules: A Reconsideration', *The Journal of International Trade and Economic Development*, vol. 15, pp. 359–73.

4.1 PRELIMINARIES

We denote by $c \geq 0$ the cost of care taken by victim and by $d \geq 0$ the cost of care taken by injurer. c will be assumed to be a strictly increasing function of care level taken by victim; and d to be a strictly increasing function of care level taken by injurer.

Let

$C = \{c \mid c$ is the cost of some feasible level of care which can be taken by victim$\}$
and
$D = \{d \mid d$ is the cost of some feasible level of care which can be taken by injurer$\}$.

We will identify $c = 0$ with victim taking no care; and $d = 0$ with injurer taking no care.

We assume:

$$0 \in C \wedge 0 \in D \tag{A1}$$

Assumption (A1) merely says that, for each party, taking no care is always a feasible option.

In view of the assumption that costs of care are strictly increasing functions of care levels it follows that:

$(\forall c_1, c_2 \in C)$ [$c_1 > c_2 \leftrightarrow$ the care level whose cost is c_1 is higher than the care level whose cost is c_2]
and
$(\forall d_1, d_2 \in D)$ [$d_1 > d_2 \leftrightarrow$ the care level whose cost is d_1 is higher than the care level whose cost is d_2].

Let Π denote the probability of occurrence of accident and $H \geq 0$ the loss in case of occurrence of accident. Both Π and H will be assumed to be functions of c and d; $\Pi = \Pi(c, d)$, $H = H(c, d)$. Let $L = \Pi H$. L is thus expected loss due to accident.

We assume:

$$(\forall c, c' \in C)\ (\forall d, d' \in D)\ [[c > c' \rightarrow \Pi(c, d) \leq \Pi(c', d)] \wedge [d > d' \rightarrow \Pi(c, d) \leq \Pi(c, d')]] \tag{A2}$$

and

$$(\forall c, c' \in C)\ (\forall d, d' \in D)\ [[c > c' \rightarrow H(c, d) \leq H(c', d)] \wedge [d > d' \rightarrow H(c, d) \leq H(c, d')]]. \tag{A3}$$

In other words, it is assumed that a larger expenditure on care by either party, given the expenditure on care by the other party, does not result in greater probability of occurrence of accident or in greater accident loss.

From (A2) and (A3) it follows that:

$(\forall c, c' \in C)\ (\forall d, d' \in D)\ [[c > c' \rightarrow L\ (c, d) \leq L\ (c', d)] \wedge [d > d' \rightarrow L\ (c, d) \leq L\ (c, d')]]$.

That is to say: a larger expenditure on care by either party, given the expenditure on care by the other party, results in lesser or equal expected accident loss.

Let $M = \{(c', d') \in C \times D \mid c' + d' + L(c', d')$ is minimum of $\{c + d + L(c, d) \mid c \in C \wedge d \in D\}\}$. Thus M is the set of all costs of care configurations (c', d') which minimize sum of costs of care (also called costs of precaution) and expected accident loss. It will be assumed that:

C, D and L are such that M is nonempty. (A4)

In this paper we will consider two different ways of defining the notion of negligence. One way of defining negligence is to compare the actual care level with a specified level of care, called the due care level. Let $\bar{d} \in D$ be the due care level. Injurer is defined to be negligent iff the care level taken by him is less than $\bar{d}$; and to be non-negligent iff the care level taken by him is greater than or equal to $\bar{d}$. This notion of negligence will be referred to as negligence(sf).

Another way to define the notion of negligence is to determine whether there exists a cost-justified untaken precaution. That is to say, whether there exists a care level taking of which would have increased costs less than the reduction in expected loss. Under this notion of negligence, injurer is defined to be negligent iff there exists a cost-justified untaken precaution; and non-negligent iff there does not exist a cost-justified untaken precaution. This notion of negligence will be referred to as negligence(up).

The rule of negligence is defined by: The injurer is liable for the entire loss iff he is negligent; and he is not liable for any part of the loss iff he is non-negligent.

It should be noted that, under the negligence rule, when the notion of negligence which is used is negligence(sf) then the accident context

is completely determined if C, D, L and $\bar{d}$ are specified; and when the notion of negligence which is used is negligence(up) then the accident context is completely determined if C, D and L are specified. $< C, D, L, \bar{d} \in D >$ will be called an application if the notion of negligence which is used is negligence(sf); and $< C, D, L >$ will be called an application if the notion of negligence which is used is negligence(up).

The expected costs of victim and injurer will be denoted by EC_1 and EC_2 respectively. Under the negligence rule, if the injurer is non-negligent then we have:

$$EC_1 = c + L(c, d) \text{ and } EC_2 = d;$$

and if the injurer is negligent then we have:

$$EC_1 = c \text{ and } EC_2 = d + L(c, d).$$

Both victim and injurer will be assumed to prefer smaller expected costs to larger expected costs and be indifferent between alternatives with equal expected costs.

If the only costs of interaction between victim and injurer are costs of care and expected accident loss, then total social costs of harmful interaction are simply the sum of them; and therefore the concept of efficiency of negligence rule can be defined as follows: The negligence rule is efficient for a given application satisfying (A1)–(A4) iff $(\forall(\bar{c}, \bar{d}) \in C \times D)[(\bar{c}, \bar{d})$ is a Nash equilibrium $\rightarrow (\bar{c}, \bar{d}) \in M]$ and $(\exists(\bar{c}, \bar{d}) \in C \times D)[(\bar{c}, \bar{d})$ is a Nash equilibrium].[4] In other words, the negligence rule is efficient for a particular application satisfying (A1)-(A4) iff (1) every Nash equilibrium is total social costs minimizing, and (2) there exists at least one Nash equilibrium. The negligence rule is defined to be efficient for a set of applications satisfying (A1)-(A4) iff it is efficient for every application belonging to the set.

If other costs[5] like the costs of using the legal system are also taken into account then the notion of efficiency would have to be modified appropriately by taking into consideration all costs of harmful interaction, not merely the costs of care and expected accident loss.

[4] Throughout this paper we consider only pure-strategy Nash equilibria.

[5] On various categories of costs of accidents see G. Calabresi (1970), *The Costs of Accidents*, New Haven: Yale University Press.

4.2 STRATEGIC CONSIDERATIONS WHEN NEGLIGENCE IS DEFINED AS SHORTFALL FROM DUE CARE

If we consider a simplified accident context in which only costs of interaction between victim and injurer are costs of care and expected accident loss then the negligence rule can be used to minimize these costs provided the due care level is set appropriately. If $\bar{d}$ is set equal to d^*, where for some c^*, $(c^*, d^*) \in M$, then under the negligence rule both victim and injurer are induced to take care levels which result in minimization of sum of costs of care and expected accident loss. In fact the following proposition holds:[6]

Proposition 4.1: *Let* $<C, D, L, \bar{d} \in D>$ *be an application satisfying Assumptions (A1)–(A4). If* $\bar{d} = d^*$, *such that* $(\exists c^* \in C)[(c^*, d^*) \in M]$ *then negligence rule is efficient for* $<C, D, L, \bar{d}>$.

Thus negligence rule is efficient for all applications satisfying (A1)–(A4) provided the due care level is set at a level, which, for an appropriately chosen victim's care level, minimizes the sum of precaution costs and expected accident loss ($c + d + L$). If the due care level is not ($c + d + L$)—minimizing then it is no longer the case that negligence rule is efficient for every application satisfying (A1)–(A4).

From the perspective of minimization of social costs it is clearly desirable that those who are in a position to undertake reduction of expected loss at a lower cost than others are the ones who take care rather than those who can do so only at a higher cost. As negligence rule under the conditions of Proposition 1 is an efficient rule it follows that if victim's costs of taking care remain the same but injurer's costs of taking care vary one would expect due care level for the injurer to tend to increase as his costs of taking various levels of care decrease as is illustrated by the following example:

Example 4.1

Suppose both victim and injurer have three alternatives: no care, moderate care, high care; and let these three alternative levels of care cost each of them 0, 1 and 2 respectively.

Thus,

[6] See Brown (1973).

$C = D = \{0, 1, 2\}$.

For $(c, d) \in C \times D$, let $L(c, d)$ be as given in the following array:

		d 0	1	2
	0	10.00	8.50	7.75
c	1	8.50	7.00	6.25
	2	7.75	6.25	5.50

Social costs $[c + d + L(c, d)]$, $(c, d) \in C \times D$, therefore are as given in the following array:

		d 0	1	2
	0	10.00	9.50	9.75
c	1	9.50	9.00	9.25
	2	9.75	9.25	9.50

Thus (1,1) is the unique social costs minimizing configuration of costs of care. If $\bar{d} = 1$ then under the negligence rule $c + d + L(c, d)$ is minimized; injurer and victim having expected costs of 1 and 8 respectively.

Suppose there is no change in costs of taking alternative levels of care by the victim; but for the injurer the cost of taking any particular level of care now is only .7 of what it was earlier. Then we have:

$C = \{0, 1, 2\}$, $D = \{0, .7, 1.4\}$.

Social costs $[c + d + L(c, d)]$, $(c, d) \in C \times D$, are as given in the following array:

		d 0	0.7	1.4
	0	10.00	9.20	9.15
c	1	9.50	8.70	8.65
	2	9.75	8.95	8.90

Thus (1,1.4) is the unique social costs minimizing configuration of costs of care. If $\bar{d} = 1.4$ then under the negligence rule $c + d + L(c, d)$ is minimized; injurer and victim having expected costs of 1.4 and 7.25 respectively.

Thus negligence rule under the conditions of Proposition 1 achieves minimization of social costs by putting greater burden on the more efficient individuals. The obverse side of this feature of negligence rule is that it tends to punish dexterity and reward incompetence. This perverse feature of the negligence rule is in some ways quite similar to the perversity of utilitarianism which would allocate most of the good things of life to gluttons with a keen sense of enjoyment rather than to those who are industrious but without highly cultivated tastes. Utilitarianism tends to reward investment of time in cultivating taste and punish investment of time in productive activities. The negligence rule tends to reward the class of injurers who are inefficient and punish those who are efficient.

In view of this perverse feature of the negligence rule if an injurer is in a position to bring about a reduction in his cost of care by expending an amount which is cost-justified, there is no guarantee that such a reduction would be undertaken. In the context of the example discussed above, improved precaution technology is resulting in reduction of social costs from 9.00 to 8.65, a net gain of 0.35. If such an improvement can be brought about at an expense of less than 0.35, then from a social perspective it would be worthwhile. But, the injurer who might be in a position to bring about such an improvement would be a loser if he does so as he will have to not only bear the expenses which would bring about such an improvement in the precaution technology in the first place; but also greater cost of care as it will go up from 1 to 1.4. Needless to say, no rational injurer would undertake such an improvement in precaution technology.

In fact, in general, there would be incentive for individuals with greater dexterity to hide their superior abilities and pretend that they also have similar abilities as the less dexterous ones. If they can successfully misrepresent their abilities of taking care they could end up paying less than otherwise. Thus, it might pay to invest time and effort in misleading the courts regarding one's ability to take care. As an illustration suppose that the actual accident context is that of Example 4.1 with improved precaution technology. If there is no misrepresentation then injurer's expected costs would be 1.4. If injurer is successful in pretending that his D is {0,1,2} rather than {0,.7,1.4} then his gain will be .4. Therefore as long as he can

successfully misrepresent by expending an amount less than .4 he would so.

Thus we see that the rule of negligence can at times provide disincentives for increasing one's efficiency by undertaking cost-justified expenses for the purpose; and incentives for expending resources for the purpose of misleading courts.

The property of negligence rule of minimizing the sum of precaution costs and expected accident loss crucially depends on the ability of courts to fix due care level on a case-by-case basis. This by itself would require expending of considerable resources. Moreover, as discussed above, parties would normally have incentives to misrepresent their abilities to take precaution for accident prevention, which would tend to further increase the costs of eliciting correct information for the purpose of fixing due care levels on a case-by-case basis.

Under the negligence rule there is one further source of strategic manipulation. This arises from the fact that taking care in one context can have beneficial implications in another context. Suppose injurer's taking care not only brings about a reduction in expected accident loss in the context of victim-injurer interaction under consideration but also in another context. Then, in calculating the optimal amount of care one should take into account not only the context under consideration but also the other context or contexts. In general, more contexts one includes, the greater would be the care which would minimize social costs. Thus, from the perspective of injurer the narrower the scope the better it is for him. From the perspective of minimization of social costs if the net benefit from taking care, taking into consideration all contexts, is positive then care should be taken. Because of limitation of knowledge, in general no injurer would be in a position to take into account all contexts. Given this limitation of knowledge, from the perspective of the society, what is desirable is that as long as the benefit of including a context exceeds the cost of required deliberation, the context should be included. As costs of deliberation are likely to be different for different injurers, what is foreseeable and what is unforeseeable would differ from injurer to injurer. As greater foreseeability would result in general in greater due care, injurers would have incentive to misrepresent their ability to foresee various contingencies.

One possible way that might be considered for eliminating strategic aspects of specification of due care level is to abandon individualized specifications and go in for a uniform specification for all injurers notwithstanding differences in costs of taking care. It is immediately clear that if there is a uniform due care for all injurers then there is no incentive for any injurer to expend resources for misrepresenting his abilities of taking care. Thus the wasteful use of resources discussed above in the context of individualized specification of due care would not arise if there is a uniform due care for all injurers. Also, if an injurer can reduce costs of care by expending resources in a cost-effective manner he would do so as the both the gains and costs would accrue to him. Another advantage of a uniform due care level is that the costs of using the legal system would tend to be smaller compared to the case when for each accident case the due care level has to be determined separately on the basis of effectiveness of the parties in taking various levels of care.

On the negative side, if there is a uniform due care level for all injurers then the property of negligence rule of minimizing the sum of precaution costs and expected accident loss no longer holds. Furthermore, it is not the case that with uniform due care there is no scope for strategic considerations. While, given the fixity of uniform due care, no individual has any incentive to misrepresent his abilities or not to undertake socially beneficial measures for increasing effectiveness of care, the set of injurers as a whole has all the incentives for strategic manipulations which have been discussed above. To see this, first we note that if negligence rule is to play some role in reducing costs of accident, then in fixing a uniform due care the average effectiveness of care-taking by injurers vis-à-vis the average effectiveness of care-taking by victims has to be taken into account. But then for the set of injurers as a whole all the points, on which individual injurers in the case of individualized due care specification found it beneficial to behave strategically, become relevant. Therefore in most instances one can expect some kind of collective or organization of injurers to emerge which would attempt to do what individuals in a setting of individualized due care can be expected to do.

4.3 STRATEGIC CONSIDERATIONS WHEN NEGLIGENCE IS DEFINED AS EXISTENCE OF A COST-JUSTIFIED UNTAKEN PRECAUTION

Before we discuss the strategic aspects of the negligence rule when the notion of negligence is defined as existence of a cost-justified untaken precaution, we note that this notion of negligence is logically completely independent of the notion of negligence as shortfall from due care where due care is defined appropriately from the perspective of minimization of social costs. This can be seen from the two examples which follow:

Example 4.2

Let $C = D = \{0,1,2\}$.
For $(c, d) \in C \times D$, let $L(c, d)$ be as given in the following array:

		d = 0	d = 1	d = 2
c	0	10.0	8.5	7.4
	1	8.5	7.0	6.1
	2	7.4	6.1	5.3

Social costs $[c + d + L(c,d)]$, $(c, d) \in C \times D$, therefore are as given in the following array:

		d = 0	d = 1	d = 2
c	0	10.0	9.5	9.4
	1	9.5	9.0	9.1
	2	9.4	9.1	9.3

Thus (1,1) is the unique social costs minimizing configuration of costs of care. If due care level for the injurer has to be set appropriately from the perspective of minimization of social costs then it has to be $\bar{d} = 1$.

Now, consider the situation when both victim and injurer are taking 0 care. Under the notion of negligence(sf) injurer is negligent because $0 < 1 = \bar{d}$. Injurer is negligent under the notion of negligence(up) as well because if injurer takes care equal to 1 expected loss will decrease by 1.5 and cost of care will increase by 1; thus $d = 1$ is a cost-justified untaken precaution.

Next consider the situation when victim is taking 0 care and injurer is taking care = 1. Under the notion of negligence(sf) injurer is non-negligent as $1 = \bar{d}$. But injurer is negligent under the notion of negligence(up) because if injurer takes care equal to 2 expected loss will decrease by 1.1 and cost of care will increase by 1; thus $d = 2$ is a cost-justified untaken precaution.[7]

Consider the situation when victim is taking 0 care and injurer is taking care = 2. Under the notion of negligence(sf) injurer is non-negligent as $2 > \bar{d}$. Injurer is also non-negligent under the notion of negligence(up) because 2 being the highest feasible level of care the question of there existing a cost-justified untaken precaution does not arise.

Example 4.3

Let $C = D = \{0, 1\}$.

For $(c, d) \in C \times D$, let $L(c, d)$ be as given in the following array:

		d	
		0	1
c	0	10.0	8.5
	1	8.6	8.0

Social costs $[c + d + L(c, d)]$, $(c, d) \in C \times D$, are as given in the following array:

		d	
		0	1
c	0	10.0	9.5
	1	9.6	10.0

Thus (0,1) is the unique social costs minimizing configuration of costs of care. If due care level for the injurer has to be set appropriately

[7] Grady (1984) proceeds under the assumption that one can infer non-existence of a cost-justified untaken precaution from the fact that injurer is taking care at a level which is greater than or equal to the social costs minimizing level. This example shows that in general such an inference cannot be made. It is only under very restrictive conditions that it would be possible to make this inference.

from the perspective of minimization of social costs then it has to be $\bar{d} = 1$.

Now, consider the situation when victim is taking care $c = 1$ and injurer is taking care $d = 0$. Under the notion of negligence(sf) injurer is negligent because $0 < 1 = \bar{d}$. Injurer is however non-negligent under the notion of negligence(up) because if injurer takes care equal to 1 expected loss will decrease by .6 only and cost of care will increase by 1; thus there does not exist any cost-justified untaken precaution.

Example 4.2 shows that it is possible for an injurer to be (i) negligent(sf) as well as negligent(up); (ii) non-negligent(sf) and negligent(up); and (iii) non-negligent(sf) and non-negligent(up). Example 4.3 shows that it is possible for an injurer to be (i) negligent(sf) but non-negligent(up). Therefore it follows that the following proposition holds:

Proposition 4.2: *The notion of negligence(sf) is logically independent of the notion of negligence(up).*

We noted earlier that if negligence is defined as shortfall from due care and due care level is set appropriately from the perspective of social costs minimization then the negligence rule is efficient for every application satisfying A(1)-A(4). The corresponding result, however, does not hold if negligence is defined as existence of a cost-justified untaken precaution, as the following example shows.

Example 4.4

Let C, D and L be as in Example 4.2.

Defining negligence as existence of a cost-justified untaken precaution, we find injurer to be negligent at (c, d), $(c, d) \in C \times D$, as given in the following array:

		d		
		0	1	2
	0	*negligent*	*negligent*	*non-negligent*
c	1	*negligent*	*non-negligent*	*non-negligent*
	2	*negligent*	*non-negligent*	*non-negligent*

Therefore the configuration of expected costs of victim and injurer $(EC_1(c, d), EC_2(c, d))$ at (c, d), $(c, d) \in C \times D$, is as given in the following array:

		d		
		0	1	2
	0	(0, 10)	(0, 9.5)	(7.4, 2)
c	1	(1, 8.5)	(8, 1)	(7.1, 2)
	2	(2, 7.4)	(8.1, 1)	(7.3, 2)

Given that the injurer is taking care $d = 1$; if victim takes care $c = 1$ then his expected costs EC_1 (1,1) will be 8, but if he takes care $c = 0$ then his expected costs EC_1 (0,1) will be 0. From this it follows that the unique social costs minimizing configuration (1,1) is not a Nash equilibrium.

In view of the above example it follows that the following proposition holds:

Proposition 4.3: *Let the notion of negligence be negligence(up). Let A be the set of all applications <C, D, L> satisfying Assumptions (A1)-(A4). Then negligence rule is not efficient with respect to A.*

While discussing strategic aspects of negligence rule, negligence being defined as shortfall from due care, we saw that the rule provided incentives for misrepresenting one's abilities to take care. When negligence is defined as existence of a cost-justified untaken precaution, the rule can be manipulated even if there is no misrepresentation of relevant facts having a bearing on the effectiveness of care by the two parties. This is because under this way of defining the idea of negligence whether one is negligent or not depends not only on one's own care level but also on the other party's care level.

This particular strategic aspect is in addition to other aspects which were discussed in the context of negligence as shortfall from due care level, where due care level is set on an individualized basis. Various ways in which the negligence rule gave rise to strategic considerations in the setting of individualized due care levels all emanated from the fact that individuals possessing greater abilities of taking care are required to undertake greater levels of care in relation to others. This particular feature is present in the notion of negligence as existence of a cost-justified untaken precaution as well.

4.4 CONCLUDING REMARKS

In this paper the negligence rule has been discussed from the perspective of strategic aspects under three different notions of negligence. If the notion of negligence is defined the way it is usually done in the law and economics literature, that is, as shortfall from due care which is set in an individualized way and at a social costs minimizing level then at least three sources of strategic manipulation arise:

(1) Individuals may not undertake cost-justified measures to increase effectiveness of their care-taking.
(2) Individuals may expend resources for the purpose of misrepresenting their care-taking abilities; such expending of resources being socially wasteful.
(3) Individuals may expend resources for socially harmful purpose of misrepresenting their abilities of foreseeing multiple risks.

When due care is set at a uniform level for all injurers, but still with a view to minimize social costs of accidents, then while individuals will have no incentive to manipulate strategically, the set of individuals as a whole will continue to have incentives to manipulate as in (1)–(3). Whether a collective organization of injurers will emerge or not will of course depend on many factors including potential gains from manipulation and transaction costs involved. When the notion of negligence is defined as existence of a cost-justified untaken precaution then in addition to (1)–(3) the possibilities of manipulation emerge also on account of the fact that negligence or otherwise of one party now may depend on what the other party does.

Although in this paper only the negligence rule has been analysed, the strategic considerations which have been discussed are relevant for a wide variety of liability rules. Indeed, the strategic considerations discussed here are relevant for every liability rule which is efficient in the same sense as the negligence rule is. If due care levels are chosen from the perspective of minimizing social costs then it can be shown that a liability rule is efficient iff its structure is such that whenever one party is negligent and the other non-negligent the entire accident

loss is borne by the negligent party.[8] As the notion of negligence under which this general characterization theorem holds is based on the idea of social costs minimization the fundamental feature because of which strategic considerations (1)–(3) become operative still holds. Thus, it is not just the negligence rule which gives rise to strategic aspects, every efficient rule does.

Another way to state the general manipulability of all efficient liability rules is the following: All the three notions of negligence which have been discussed are such that they give rise to strategic considerations. All the three notions incorporate in some way the objective of minimization of social costs of accidents. Minimization of social costs requires allocating care-taking in such a way that those who are better at it do more of it compared to the others. To induce more able persons to take more care, the negligence standards must be chosen in such a way that more able persons have to take greater care in order to escape from being adjudged negligent compared to the others. Thus notions of negligence having the perverse implication of punishing ability and dexterity is a direct consequence of attempting to use the idea for minimization of social costs. This perverse implication, however, provides inappropriate signals to rational individuals giving rise to possibilities of strategic behaviour which will work in the opposite direction, that is, against the objective of minimization of social costs. While in the law and economics literature question of efficiency of liability rules has been analysed extensively and the economic logic of notions of negligence studied in depth, the strategic aspects associated with these notions of negligence have not received the requisite attention. There is need to look at these strategic aspects more closely as from the preliminary analysis carried out in this paper it appears to be the case that achieving the objective of minimization of social costs by the use of liability rules may be much more difficult than has been thought to be the case hitherto.

REFERENCES

Brown, J.P. (1973), 'Toward an Economic Theory of Liability', *Journal of Legal Studies*, vol. 2, pp. 323–50.

[8] See S.K. Jain and R. Singh (2002), 'Efficient Liability Rules: Complete Characterization', *Journal of Economics (Zeitschrift für Nationalökonomie)*, vol. 75, pp. 105–24.

Calabresi, G. (1970), *The Costs of Accidents*, New Haven: Yale University Press.

Cooter, R.D. and T.S. Ulen (2003), *Law and Economics*, 4th edn, New York: Addison-Wesley.

Grady, M.F. (1983), 'A New Positive Theory of Negligence', *Yale Law Journal*, vol. 92, pp. 799–829.

—— (1984), 'Proximate Cause and the Law of Negligence', *Iowa Law Review*, vol. 69, pp. 363–449.

—— (1989), 'Untaken Precautions', *Journal of Legal Studies*, vol. 18, pp. 139–56.

Jain, S.K. (2006), 'Efficiency of Liability Rules: A Reconsideration', *The Journal of International Trade and Economic Development*, vol. 15, pp. 359–73.

Jain, S.K. and R. Singh (2002), 'Efficient Liability Rules: Complete Characterization', *Journal of Economics (Zeitschrift für Nationalökonomie)*, vol. 75, pp. 105–24.

Landes, W.M. and R.A. Posner (1987), *The Economic Structure of Tort Law*, Cambridge, MA: Harvard University Press.

Posner, R.A. (1972), 'A Theory of Negligence', *Journal of Legal Studies*, vol. 1, pp. 28–96.

—— (2007), *Economic Analysis of Law*, 7th edn, New York: Wolters Kluwer Law & Business.

Shavell, S. (1987), *Economic Analysis of Accident Law*, Cambridge, MA: Harvard University Press.

5

SUIT AND SETTLEMENT UNDER ASYMMETRIC INFORMATION

The Case of Bhopal Gas Disaster

P.G. Babu*

One of history's worst industrial disaster at the Bhopal's pesticide plant of Union Carbide India Ltd (UCIL) in India in 1984, and the subsequent litigation battle between the Government of India and the Union Carbide Corporation (UCC) provides an important opportunity to explore the policy relevance of the theories of bargaining and informational asymmetry.

This case began at Kenan, J's court at New York in 1985 with multi-billion dollar claims and ended in the Supreme court of India in 1991 with a 540 million US dollars out-of-court settlement. It raised many puzzling questions. First, one can question the Government of India's action of taking over the legal powers of the victims: Why should the government restrict the victim's right to sue for damages? The second issue concerns the Indian government's decision to file the suit in an American court. Why should a sovereign state denigrate its own legal system in a foreign court? The third question pertains to the decision of the lawyers representing private interests to accept 300 million US dollars as compensation. The same lawyers were claiming around 15 billion US dollars just a year earlier. Why should they do such a drastic volte face? Most importantly, the Government of India and

* The author is grateful to Richard Day, Ajit Mishra, and Anindya Sen for their comments on an earlier version. The usual disclaimer applies.

UCC agreed to 540 million US dollars after a long delay; why did not they do so immediately? In other words, what were the reasons for the bargaining delay? We will try to analyse these questions and attempt to answer them using the results from information economics and bargaining theory.

5.1 THEORETICAL BACKGROUND

One of the fundamental hypotheses of economics of law as laid out by Coase[1] is that voluntary bargaining mitigates efficiency. In case a dispute arises between a potential plaintiff and a defendant, the parties have opportunities to settle voluntarily out-of-court instead of engaging in the costly process of litigation. But, in reality, many cases do go to court even though the costs associated with such litigations are enormous.[2] If it is possible to settle out-of-court efficiently, why should litigation occur as it did in the case under consideration.

Numerous attempts have been made to resolve this puzzle. Landes,[3] Gould,[4] Posner,[5] and Shavell[6] argue that differing views and assessments of the parties involved about the outcome of the trial might be a major reason behind litigation attempts. Bebchuk,[7] P'ng,[8] Salant and Rest,[9]

[1] R. Coase (1960), 'The Problem of Social Cost', *Journal of Law and Economics*, vol. 3, pp. 1–44.

[2] R.D. Cooter and D.L. Rubinfeld (1989), 'Economic Analysis of Legal Disputes and Their Resolution', *Journal of Economic Literature*, vol. 27, pp. 1067–97.

[3] W. Landes (1971), 'An Economic Analysis of the Court', *Journal of Law and Economics*, vol. 14, pp. 61–107.

[4] J. Gould (1973), 'The Economics of Legal Conflict', *Journal of Legal Studies*, vol. 2, pp. 279–300.

[5] R.A. Posner (1973), 'An Economic Approach to Legal Procedure and Judicial Administration', *Journal of Legal Studies*, vol. 2. pp. 399–458.

[6] S. Shavell (1982), 'Suit, Settlement and Trial: A Theoretical Analysis Under Alternative Methods for the Allocation of Legal Costs', *Journal of Legal Studies*, vol. 11, pp. 55–81.

[7] L.A. Bebchuk (1984), 'Litigation and Settlement under Imperfect Information', *Rand Journal of Economics*, vol. 15, pp. 404–15.

[8] I.P.L. P'ng (1983), 'Strategic Behavior in Suit, Settlement, and Trial', *Bell Journal of Economics and Management*, vol. 14, pp. 539–50.

[9] S.W. Salant and G. Rest (1982), *Litigation of questioned settlement claims: A Bayesian Nash equilibrium approach*, Discussion Paper p6809, Rand Corporation.

Salant,[10] and Samuelson[11] consider the strategic motives involved in litigation by explicitly modelling litigation as an extensive form game.

As opposed to the widespread recourse to litigation, there is also some empirical evidence that in many situations involving patent infringement, price discrimination, predation, fraud, etc., more than 80(1987). The fact that cases settle in the shadow of the law does not mean that the legal system is irrelevant. The formal legal system substantially influences the behaviour and precautionary measures taken by the parties concerned as well as the settlements reached outside the courts. Therefore, a fundamental question before economic theorists is to explore the links between law and information economics.

5.2 THE CASE HISTORY

5.2.1 The Company[12]

In 1966, Union Carbide India Ltd (UCIL) proposed to set up a modern pesticide plant to capture the market potential during India's 'Green revolution' phase and the Government accepted the proposal. Initially the pilot plant was started in Trombay in Maharashtra; but, in 1968, the company unilaterally decided to shift the plant to Bhopal in the state of Madhya Pradesh. Initially it was importing Methyl Iso Cyanate (MIC) to produce Sevin and Temik, the brand of pesticides it produced. In 1980, it started its own MIC plant with the government's permission. The MIC plant was riddled with technical problems right from the construction stage. Prior to the major accident in 1984, there were two other accidents in which one worker died and many others were to be treated for health problems arising from MIC leakages. The employee union had even threatened to go on strike demanding enhanced plant safety measures.

[10] S.W. Salant (1984), *Litigation of settlement demands questioned by Bayesian defendants*, Social Science Working Paper 516, California Institute of Technology.

[11] W. Samuelson (1983), *Negotiation v. Litigation*, Discussion paper, School of Management, Boston University.

[12] The case history is based on the following sources: J. Cassels (1993), *The Uncertain Promise of Law*, University of Toronto Press, Toronto, Buffalo; S. Hazarika, 1988, *Bhopal: The Lessons of a Tragedy*, Penguin Books, New Delhi; and the Case of Union Carbide Corporation Gas Plant Disaster at Bhopal, India, 634F. Supp. 842 Southerns District Court of New York, 1986.

5.2.2 The Government[13]

The state government of Madhya Pradesh was happy to have the MIC plant at Bhopal as a showpiece for the industrial development of the state. Its Town and Country Planning Board classified the Bhopal plant of UCIL as a 'general industry' instead of 'hazardous industry'. This classification allowed the company to set its plant very near the then city limits. Later, the state government allowed slums and residential colonies to develop around the plant and legalized it for political purposes. Also, the state Inspectorate of Factories was highly understaffed; most of its employees were mechanical engineers with little or no knowledge of chemical processes. Despite a major leak of MIC in the plant, the inspectorate and the industry minister had given a clean chit to the factory.

5.2.3 The Disaster[14]

On the fateful night of 2 December 1984, water entered one of the MIC storage tanks and started reacting with it. The temperature within the tank soared over 400°C, which led to the shattering of the concrete tank and the eventual escape of the poisonous gas. The vent gas scrubbers which were supposed to neutralize the gas by sprinkling caustic soda were not working.

The gas, unhindered, spread over the nearby slum and other parts of the city in the direction of the wind, killing about 2,850 people and impairing the health of more than one hundred thousand persons. As per the Indian Council of Medical Research (ICMR) investigation, of the total of 2,566 women who were pregnant during the disaster, 355 had spontaneous abortions and 56 had still-born babies. There was a high incidence of infant mortalities, widespread deformities, and disturbances of the central nervous system among the affected human populace as well as livestock. Approximately 25,000 of the affected persons are still suffering from the long-term effects, which includes mental health problems.

The state government's disaster management potential was highly unsatisfactory. In fact, it could not find the exact number of victims, living and dead, until late 1987, thirty four months after the disaster.

[13] Ibid.

[14] Ibid.

Also, it could not come up with adequate relief and reimbursement measures immediately after the accident.

5.2.4 The Litigation[15]

Immediately after the disaster, thousands of litigation suits demanding billions of US dollars in damages were filed in the various US courts by private parties. They sought damages and punitive compensations under US tort law for the loss of human lives and the misery of those who survived the disaster. The US federal judicial panel consolidated all these cases and brought it before John F. Kenan, J of the federal district court of the south district of New York.

Numerous US lawyers signed contingent fee contracts with the survivors in order to represent them in US courts. These contracts allowed the lawyers to take a certain percentage of the compensation paid, contingent on their winning the case, which is a common feature in US tort cases. In order to prevent this practice, the Indian government issued an act by which it declared itself to be the sole representative of the gas victims. However, since such an Indian law would not hold good in foreign courts, through a separate clause the government allowed private representation in foreign courts at the cost of the compensation seeker.

Finally, on 8 April 1985, Michael Ciresi of the law firm 'Robins and Zelle' filed a suit in Kenan's court on behalf of the Indian government. All other private interests were represented by Stanley Chesley and Lee Bailey, who filed a separate suit on 26 June 1985. These suits were significant in one respect, that they did not mention any specific compensation figure sought, on the ground that more surveys and scientific studies would have to be undertaken.

Kenan ordered UCC to pay an interim relief of 5 to 10 million US dollars on humanitarian ground. UCC responded with a 5 million US dollar offer which was refused by the Indian government on the grounds that the judge imposed too many conditions on disbursement. In the meantime, the president of UCC announced that his company did not have the financial capacity to meet the kind of claims made,

[15] Case of Union Carbide Corporation Gas Plant Disaster; United States Court of Appeals, Second Circuit, 809 F. 2d 195; and *United Carbide Corporation* v. *Union of India and Others*, AIR Supreme Court 273.

after the negotiations with the Indian government failed. By the end of June 1985, for the first time, UCC presented its settlement offer of 100 million US dollars. After this offer was refused by the Government of India, UCC filed its affidavit, in which it tried to disassociate from its Indian subsidiary. It contended that the case should be tried in Indian courts under Indian regulations.

The Government of India pointed out that its legal system was insufficient due to the undeveloped nature of its tort law.[16] Marc Galanter who appeared as a witness pointed out that the discovery procedures in Indian courts were very weak. Kenan initiated discovery proceedings in the month of October, 1985. At the suggestion of Kenan, both the parties again negotiated 'under the shadow of law'. UCC raised its settlement offer from 100 million to 300 million US dollars. Again, the Indian government refused. Immediately after this refusal, UCC approached the lawyers representing private interests and made the same offer. This offer was accepted by the private lawyers subject to a full agreement on 19 March 1986. Once this information became public, Kenan stopped all efforts for an out-of-court settlement.

On 12 May 1986, Kenan delivered his judgment in which he sent the case back to Indian courts under the condition that UCC, and not its subsidiary, is bound by Indian court verdict and UCC should subject itself to American discovery and civil procedures in Indian courts. UCC opposed the second condition and appealed. This appeal was granted by US court of appeals for the second circuit and later upheld by US supreme court.

The case moved to the court of J.S. Patel, J of the Bhopal district courts. In November 1986, the Indian government for the first time announced its compensation demand for 3.1 billion US dollars. UCC had to fight a takeover bid in the mean time by selling a considerable portion of its fixed assets. Alarmed by this, the Indian government moved the court. Patel ordered the company to pledge 3 billion US dollars which was accepted by UCC without acknowledging liability for the disaster. Later, it turned out that Patel himself was a claimant, which raised issues of mistrial. Finally, in late 1988, under the

[16] J. Cassels (1993).

instructions of the Supreme Court of India, both the parties negotiated once again. UCC's new offer of 540 million dollars was accepted by the Indian government. This settlement was contested in the Supreme Court of India which in October 1991, upheld the settlement amount but held UCC liable for criminal negligence.

5.3 ANALYSIS

A careful perusal of the above litigation between UCC and the Government of India and private parties in the courts of United States and India raises many issues and questions. First, one can question the Government of India's decision of taking over the legal powers of the victims: Why should the government restrict the victims' right to sue for damages? The second issue concerns the Indian government's decision to file the suit in an American court: Why should a sovereign state denigrate its own legal system in a foreign court? The third question pertains to the decision of the lawyers representing the private interests to accept 300 million US dollars as compensation. The same lawyers were claiming around 15 billion US dollars just a year earlier. Why should they do such a drastic volte face? Most importantly, the Government of India and UCC agreed to 540 million US dollars after a long delay; why did not they do it immediately? In other words, what were the reasons for the bargaining delay? We will try to analyse these questions and attempt to answer them using the results from information economics and bargaining theory.

5.3.1 Abrogation of Legal Rights

The decision of the Indian government to be the sole representative of the Bhopal gas victims was one of the most controversial issues in this legal battle. We argue that this decision was correct in the light of bargaining strength, moral hazard, and unity problems.

Consider the settlement bargaining game between a plaintiff and the defendant, where they bargain over the amount of compensation, C. It would suffice to restrict ourselves to the bargaining game with complete information. The plaintiff and the defendant alternate in making offers over time $t = 0, 1, 2, \ldots$ It is assumed that during the odd-numbered periods, the plaintiff makes a demand for the compensation amount which the defendant can accept or reject. If

the defendant accepts, the settlement bargaining ends whereby the defendant transfers the agreed compensation to the plaintiff. If it is rejected, the plaintiff can decide whether to sue.

During the even numbered time periods, the defendant offers a compensation which the plaintiff may accept or reject. Acceptance ends the game and rejection leads to either filing a suit or not doing so by the plaintiff. The discount factors of the plaintiff and the defendant are given by $\delta_P \in (0, 1)$ and $\delta_D \in (0, 1)$ respectively for $t = 0, 1, 2, \ldots$ Supposing the plaintiff can precommit to file a suit whenever there is disagreement, we get the following result due to Rubinstein.[17]

Proposition 5.1 (Rubinstein)

If the plaintiff can precommit to file a suit whenever there is disagreement, then there is an unique subgame perfect equilibrium for the settlement bargaining game, whereby the settlement is reached in the first period itself. The share of the plaintiff equals $(1-\delta_D)C/(1-\delta_P\delta_D)$ if the bargain starts in an odd time period and equals $\delta_P(1-\delta_D)C/1-\delta_P\delta_D$ if it begins at an even time period.

This result emphasizes the importance of the 'impatience level', that is, the discount factor, in a settlement bargain. Leaving aside the issue of the first mover advantage by assuming that the plaintiff moves first, he will get C if $\delta_P \to 1$ (highly patient) and $(1-\delta_D)C$ if $\delta_D \to 0$ (highly impatient).

In the case of Bhopal disaster, most of the victims belonged to the slums around the UCIL factory. They were mostly daily wage workers whose average yearly income was around Rs 4,000. For them, a promised income in the future is worth far less than that of today's. In other words, they are highly impatient as compared to UCC. Hence, they are likely to accept whatever settlement offer is made, immediately. Therefore, if UCC had to negotiate directly with the victims, it would have gotten away with a large share of the compensation under negotiation. Also, the probability of an early court verdict, even in US courts, was very low (which is the case in the Indian courts). For example, the litigation between Pennzoil and Texaco took four years to

[17] A. Rubinstein (1982), 'Perfect Equilibrium in a Bargaining Model', *Econometrica*, vol. 50, pp. 97–109.

settle.[18] Taking these points into account, one can safely conclude that the bargaining power of the victims was very small. The government, on the other hand, is capable of withstanding long bargaining delays due to its higher patience level.

Apart from the bargaining power, the Government of India was alarmed at the way the US lawyers were signing the contingent contracts. These contracts provided the lawyers approximately 50 per cent of the total expected damage compensation. Such a huge contingent fee and agency problems that went with the contingent contracts provided the second major reason for the government's action.

Finally, the large number of victims would have enabled UCC to offer discriminating settlements. In a direct settlement game between UCC and the victims, it would have been impossible to find coalition proof the Nash equilibrium; that is, some sub-coalition of victims would find it beneficial to break the coalition. Such behaviour would have helped UCC to minimize its settlement amount.

To prove the above aspects, let us consider the following example adapted from Bernheim *et al.*[19] In this three player game, player 1 has two strategies that is, a_1 and a_2; player 2's strategies are given by b_1 and b_2. Player 3 chooses one of the two game matrices denoted by c_1 and c_2. The entries in the game matrix stand for the payoffs of the players; for example, (1, 1, −5) will be the resultant payoff of the players 1,2 and 3 if they choose a_1, b_1, and c_1 respectively as their strategies. Other entries can be interpreted similarly.

Example:[20]

	c1	
	b1	b2
a1	1, 1, −5	−5, −5, 0
a2	−5, −5, 0	0, 0, 10

	c1	
	b1	b2
a1	−1, −1, 5	−5, −5, 0
a2	−5, −5, 0	−2, −2, 0

For this game (a_2, b_2, c_1) and (a_1, b_1, c_2) form the set of Nash equilibria. By the principle of Pareto dominance, one can choose (a_2, b_2, c_1) as an

[18] Cooter and Rubinfeld (1989).

[19] D. Bernheim, B. Peleg, and M. Whinston (1987), 'Coalition-proof Nash Equilibria, 1, Concepts', *Journal of Economics Theory*, vol. 42, pp. 1–12.

[20] Ibid.

equilibrium. But, players 1 and 2 can form a coalition together and choose to play (a_1, b_1). This can be avoided if the player 3 plays c_2. The point to note here is that left to themselves, the victims or their lawyers would have broken the unity achieved by the Government's action. This is the third major reason supporting the Government's decision.

5.3.2 The Choice of the Foreign Court

The Government of India's decision to try the case in an US court stems from the following issues: (i) the difference between two legal standards: strict liability and negligence, and (ii) pretrial discovery rules prevalent in the US courts.

In trying the tort cases, courts resort to either the strict liability standard or the negligence standard. Under the strict liability standard, once the plaintiff proves that he is harmed by the defendant's action, the latter is held liable irrespective of his level of care in preventing the accidents. On the other hand, under negligence standard, the defendant is held liable only if he is proved to fall short of the legally defined optimum level of care. The general practice is to use negligence rule when there exists a socially defined optimum care level and perfect information about the care chosen by the defendant, and strict liability in cases where the defendant's care remains private information and one cannot define such optimal care levels.

To use the negligence standard, one has to define the legal definition of optimum or reasonable care. This definition was developed by US federal judge Hand and is popularly known as the 'Hand rule'.[21] Let b be the benefit to the injurer (defendant) from undertaking an action and H be the harm done to the injured (plaintiff) by this action with probability p. Then, if $b < p.H$, the action of the injurer is wrong and he will be held liable. But, the Hand rule standard will be violated whenever the compensation is too small or the probability of enforcement is too low. This violation would lead to either underprotection or overprotection by the injurer.

Under the negligence standard, the onus is on the plaintiff to prove that the defendant has fallen short of the optimum standard for the care levels.

[21] R.D. Cooter (1991), 'Economic Theories of Legal Liability', *Journal of Economic Perspectives*, vol. 5, pp. 11–30.

Under the conditions of asymmetric information, where the plaintiff does not know the level of care taken by the defendant, the litigation expenses will be enormous. This can be seen from the following simple exercise.

Consider the case of strict liability. Here, it is sufficient for the plaintiff to prove that he has suffered from the defendant's action. Let the probability of the court finding that the plaintiff has suffered due to the defendant's action when the plaintiff spends P on proving that he has indeed suffered be $p(P)$. In case the plaintiff wins, he gets C as compensation in addition to the court costs incurred by him, viz., P. Now, the plaintiff wishes to maximize his expected payoff, which is given by $p(P)(C + P) + (1 - p(P))(-P - D)$ where D stands for the defendant's cost of litigation. Here, we assume that the plaintiff bears the court expenditures of the defendant in case the latter wins.

For the case of negligence standard, the plaintiff has to prove that the defendant has violated the optimal care standard in addition to proving the harm done to him. Let $q(P)$ be the probability of the court finding the defendant guilty when the plaintiff spends P. Here, the plaintiff's expected payoff is given by $p(P)q(P)(C + P) +(1 - q(P))(-P - D)$. The plaintiff obviously wishes to maximize his expected payoff. By solving the first order equations of both the cases for P, one can easily see that the litigation costs associated with the negligence rule is more. This has serious implications. The increase in the plaintiff's litigation costs reduces the minimum possible amount which he can get in a settlement game, viz., $(C - P)$. Also, it reduces the incentive of the plaintiff to file the suit and the level of care chosen by the defendant in carrying out his action. Proving that UCC had not met the safety standards is an impossible task for two reasons. Firstly, the safety standards were neither prescribed by the inspectorate of factories nor enforced. The second, and the more important reason is that the discovery procedures of Indian courts are ineffective; parties involved can simply refuse to show the documents under various pretexts. Pre-trial discovery procedures of the US courts, on the other hand, allow the parties to get evidence (such as inspection of the documents, interrogations, etc.,) about their opponents before the actual trial, which otherwise will not be disclosed to them. The right of pre-trial discovery helps one to know more about the private information held

by the opponent. Hence, the strict liability standard of the American legal system was preferred by the Indian government to that of the negligence standard of the Indian courts. This experience has an important lesson for the Indian legal system. As neither the judges nor the regulatory agencies are in a position to set optimal care levels, the Indian courts should shift to the strict liability standard in the cases of ultra-hazardous activities. Such a shift will be socially optimal as the parties are motivated to balance the true cost of reducing risky activities against the expected savings in losses caused.

5.3.3 Agency Problems

An important instance of agency problem occurs in the process of settlement bargaining. In the bargaining process per se, the parties do not involve themselves directly; they are represented by their lawyers. Making sure that the objectives of the lawyers coincide with those of the parties by writing efficient contracts is an agency problem embedded in an overall bargaining model. The embedding can be formally viewed as the first stage of a two-stage game where the parties hire agents who in turn play the bargaining game in the second stage.

One of the crucial points of the Bhopal case is the acceptance of 300 million US dollars as settlement by the US lawyers representing the private interests (of victims). Immediately after the testimony of Palkhivala, the famous Indian lawyer and diplomat who defended the efficiency and competence of the Indian judicial system, the chances of the case returning to Indian courts went up. This implied that the private lawyers' expected fee of 50 per cent of the compensation would turn out to be zero. As soon as the lawyers revised their probabilities of a favourable verdict, they realized that 150 million dollars was perhaps the best they could get. These revised beliefs of the lawyers led to their immediate acceptance of the UCC offer.

Here, one can clearly see the conflict of objectives between the victims and the lawyers representing them. The actions undertaken by the lawyers are not known to the victims (hidden action); also, new information (that is, revised probability of favourable outcome) was becoming available only to the lawyers (hidden information). When there is hidden action and hidden information, an optimal contract should have some fixed component. Such a fixed component can be

in the form of a flat monthly fee. The variable component provides an incentive to the lawyers to put in more trial effort, while the fixed component protects them from unfavourable outcomes, such as a negative verdict.

In the Bhopal case, the victims were forced into an outcome contingent contract with only a variable component as the victims, by and large, were unable to enter into any other type of contracts. Since this did not protect the lawyers from the adverse situation, they resorted to maximize their own payoff without taking into account their clients' interests. Under the Indian system, the lawyers generally get a fixed fee and hence, there is no incentive for them to put in more trial effort apart from reputation effects. On the other hand, the US system of pure contingent fee widens the divergence of goals. Hence, under informational asymmetry shifting to a contract with both fixed and variable components will increase the second best social welfare.

This case also provides another instance of agency problem. The government of India's interests in delaying the settlement so as to obtain a higher settlement might have been far removed from those of the Bhopal victims, who might have preferred an immediate, lower settlement given their high discount factor. As soon as the case was returned to India, the government should have pursued serious negotiations, taking into account the limitations of Indian tort law. Also, the government signalled a wrong image to Kenan by refusing to accept the humanitarian aid of 5 million dollars on the basis of conditions for disbursement. This signalling effect was strengthened when the Government's relief measures were found to be wanting by Kenan. Most importantly, the government could not even produce a final list of victims even by the year 1987. These signalling effects might have played a major role in the verdict given by Kenan.

5.3.4 Bargaining Delay

Apart from the agency problems detailed above, signalling effects were the major reasons for the delay in settlement. By acting tough, the Government of India signalled to UCC that its chances of winning the case were high. UCC retaliated by a low offer of 100 million dollars and the threat that its financial position was bad. Also, the Indian government might have deliberately created uncertainty by not

mentioning the compensation figure it sought; so as to strategically influence the opponent's beliefs about winning.

By rejecting the UCC's offers, and by refusing to name the compensation figure, the Indian government escaped the Coase conjecture.[22] According to the conjecture, the readiness to make offers frequently encourages the other party to anticipate and wait for even lower offers that may produce a quick settlement. This was unfavourable to the Indian government. By prolonging the confrontational posture, the Indian government (acting on behalf of the victims) could gain a larger share of the amount under bargain. This strategy of building reputation was a dominant strategy in the US court. But, following the same strategy in Indian courts proved to be counter-productive. The government could have followed a 'skimming' strategy in the Indian court according to which it could have started from a higher demand for compensation and successively reduced it to lower demands. However, such a strategy was not followed and the government lost the chance of getting a higher compensation and also had to incur costly delays.

We can easily demonstrate that even in a bargaining game with complete information, costly bargaining delays are possible. Recall the notations and the basic model of bargaining from section 5.3.1, in which the minimum amount the plaintiff will accept is the actual harm suffered by him expressed in monetary units. Let this minimum amount be C_0. For the plaintiff to sue and continue the trial for T time periods, he should prefer receiving C after T periods instead of receiving C_0 in the first period itself. For this to happen, $\delta^T{}_P C > C_0$.

Let S be the true capability of the defendant to pay for the harm commited to the plaintiff. If the defendant attempts to reach immediate settlement by offering $s_1 + \varepsilon$ where

$$s_1 = \frac{\delta_P(1-\delta_D)C}{1-\delta_P\delta_D}, \varepsilon > 0,$$

it is not optimal for the plaintiff to reject this offer. This is so because the plaintiff will get $b_1 = (1-\delta_D)C/1-\delta_P\delta_D$ in the next period and $s_1 + \varepsilon \geq \delta_P((b_1)$. Hence, for the defendant to prefer to wait for T periods, $S - s_1 \leq \delta_D^{T-1}(S - \hat{C})$.

[22] R. Coase (1972), 'Durability and Monopoly', *Journal of Law and Economics*, vol. 15, pp. 143–9.

Combining both the conditions for the plaintiff and the defendant to wait for T periods, we get,

$$(1-\delta_D^{(1-T)})\,S+\delta_D^{(1-T)}\,s_1 \geq \hat{C} \geq \delta_D^{(-T)}\,C_o.$$

Now, we can state the following proposition which asserts that bargaining delay is possible even in a game of complete information.

Proposition 5.2 (Fernandez and Glazer)[23]

If C is such that $(1-\delta_D^{(1-T)})\,S+\delta_D^{(1-T)}\,s_1 \geq \hat{C} \geq \delta_P^{(-T)}\,C_0$, *then there is a subgame perfect equilibrium in the play which leads to a prolonged trial to T periods, followed by the settlement* $\hat{C}$.

This result is the same as Theorem 2 of Fernandez and Glazer.[24] Thus, one can hope to explain the huge bargaining delay of seven years for the settlement of the Bhopal case.

5.3.5 Bargaining Inefficiency

After a long delay of seven years, the Government of India agreed to an out-of-court settlement of 540 million dollars. Many considered this settlement to be inefficient as the government was asking for 3 billion dollars in 1987. One can explain this bargaining inefficiency based on the informational asymmetry aspects. We prove this result using a one-shot bargaining model where the players use minimax loss strategies. The advantage in using the minimax strategies is that the bargaining inefficiency can be reduced, although not completely.

The plaintiff and the defendant do not know each other's true valuations; the plaintiff does not know the true capability of payment by the defendant and the latter does not know the former's actual loss. But, the plaintiff can make an estimate of the upper bound of the defendant's capability to compensate (namely, S_{max}) and the defendant can make an estimate of the lower bound of the plaintiff's loss, given by C_{min}.

The one shot game proceeds as follows. Both the parties submit their bids (viz., c for the plaintiff and s for the defendant) to the judge and he sets the compensation at $K = (c+s)/2$ if $s \geq c$. In case, $s < c$,

[23] R. Fernandez and J. Glazer (1991), 'Striking for a Bargain Between Two Completely Informed Agents', *American Economic Review*, vol. 81, pp. 240–52.

[24] Ibid.

the settlement effort fails and the plaintiff initiates the trial effort. Let us assume that $S_{max} \geq C_{min}$, $s \leq S_{max}$ and $c \geq C_{min}$. The first condition ensures that bargaining will not fail; the other conditions state that the defendant will never bid anything above his capability and the plaintiff will not accept any offer if it is less than his actual loss.

Suppose $s \geq c$ and hence the settlement occurs. Now, we can define the defendant's gain from the settlement game as

$$\Pi_D = \begin{cases} S - \frac{(c+s)}{2} & \text{if } s \geq c \\ 0 & \text{if } s < c \end{cases}$$

The defendant would have obtained the maximum gain if he bid c. In this case, we can write his maximum gain obtainable as

$$\Pi_D^{*} \begin{cases} S - c \text{ if } S \geq c \\ 0 \quad \text{if } S < c \end{cases}$$

The defendant's loss function can be defined as $L_D = \Pi_D^{*} - \Pi_D$, which is given by,

$$L_D = \begin{cases} \frac{(s-c)}{2} & \text{if } s \geq c \\ S - c & \text{if } s < c \leq S \\ 0 & \text{if } c < S \end{cases}$$

The defendant suffers maximum loss when $c = C_{min}$ (when the settlement occurs) and $c = s$ (when the settlement fails). Replacing these identities in the above loss function we get the maximum loss of the defendant. That is,

$$\text{Maximum } L_D = \max \{\frac{(s - C_{min})}{2}, S - s, 0\}.$$

But, we are interested in the minimum of the maximum loss. That happens when

$$\frac{(s - C_{min})}{2} = S - s.$$

From this, we can get the minimax loss bid of the defendant,

$$s = (\frac{2}{3})\, S + (\frac{1}{3})\, C_{min}.$$

Similarly, we can calculate the plaintiff's minimax loss bid, which is given by

$$c = (\frac{2}{3})\, C + (\frac{1}{3})\, S_{max}.$$

If both the parties present their minimax loss bids to the judge, the compensation amount will be,

$$K = (\frac{1}{3})(S + C) + (\frac{1}{6})(S_{max} + C_{min}).$$

For a settlement to be efficient, $s - c \geq 0$. Now in the case of settlement with incomplete information, by subtracting the plaintiff's minimax bid from that of the defendant, we get,

$$S - C \geq (\frac{1}{2})(S_{max} - C_{min}) > 0.$$

Hence, the final settlement is inefficient. In the process, we have also proved that Coase's result that 'bargaining leads to efficient settlement'[25] does not hold good in the context of a settlement bargaining game with incomplete information.

5.4 SOME OBSERVATIONS

Several observations can be made from the case between the Government of India and UCC. The most important is that bargaining under the shadow of the court involves distributional and efficiency questions that are at odds with each other. The Bhopal case provides a situation where rational choices by both the parties led to sub-optimal outcome. Both the parties could have benefited by openness, communication, and joint problem-solving. But they resorted to conflict and ended up in a prisoner's dilemma situation; that is, both parties engaged in tactics that were value reducing and created unnecessary costs. Thus, the rational choices by both parties led to suboptimal outcomes for both.

A second observation in this case relates to agency problems. For researchers concerned with law and economics, it is important to study the incentives created by these agency problems and the efficiency costs they may generate in resolving the conflict. Revisions in law systems such as the introduction of contingent claim contracts with fixed and variable parts and strict liability legislation could be given due consideration.

Another major issue raised by the Bhopal gas disaster is the relevance of the regulatory agency (the Inspectorate of Factories) in mitigating such accidents. Most of the employees of the inspectorate

[25] Coase (1960).

were mechanical engineers with no knowledge of chemical processes. In other words, they had imperfect information about the relevant care levels. In such cases, it is advisable to complement the safety regulation with strict liability standards. Neither the court nor the regulatory agency on their own, would be able to induce the socially optimal level of care.

REFERENCES

L.A. Bebchuk (1984), 'Litigation and Settlement under Imperfect Information', *Rand Journal of Economics*, vol. 15, pp. 404–15.

D. Bernheim, B. Peleg, and M. Whinston, 'Coalition-proof Nash Equilibria, 1, Concepts', *Journal of Economics Theory*, vol. 42, pp. 1–12.

R. Coase (1960), 'The Problem of Social Cost', *Journal of Law and Economics*, vol. 3, pp. 1–44.

—— (1972), 'Durability and Monopoly', *Journal of Law and Economics*, vol. 15, pp. 143–49.

R.D. Cooter (1991), 'Economic Theories of Legal Liability', *Journal of Economic Perspectives*, vol. 5, pp. 11–30.

R.D. Cooter and D.L. Rubinfeld (1989), 'Economic Analysis of Legal Disputes and Their Resolution', *Journal of Economic Literature*, vol. 27, pp. 1067–97.

R. Fernandez and J. Glazer (1991), 'Striking for a Bargain between Two Completely Informed Agents', *American Economic Review*, vol. 81, pp. 240–52.

J. Gould (1973), 'The Economics of Legal Conflict', *Journal of Legal Studies*, vol. 2, pp. 279–300.

S. Hazarika (1987), *Bhopal: The Lessons of a Tragedy*, New Delhi: Penguin Books.

A.B. Krueger (1991), 'The Evolution of Unjust-dismissal Legislation in the United States', *Industrial and Labor Relations Review*, vol. 44, pp. 644–60.

W. Landes (1971), 'An Economic Analysis of the Court', *Journal of Law and Economics*, vol. 14, pp. 61–107.

R. Mnookin, and R. Wilson (1989), 'Rational Bargaining and Market Efficiency: Understanding *Pennzoil* v. *Texaco*', *Virginia Law Review*, vol. 75, pp. 295–334.

J.M. Perloff and D.L. Rubinfeld (1987), 'Settlements in Private Antitrust Litigation', in S. Salop and L. White (eds), *Private Antitrust Litigation*, Cambridge, Mass: MIT Press, pp. 149–84.

I.P.L. P'ng (1983), 'Strategic Behavior in Suit, Settlement, and Trial', *Bell Journal of Economics and Management*, vol. 14, pp. 539–50.

—— (1987), 'Litigation, Liability, and Incentives for Care', *Journal of Public Economics*, vol. 34, pp. 61–86.

R.A. Posner (1973), 'An Economic Approach to Legal Procedure and Judicial Administration', *Journal of Legal Studies*, vol. 2. pp. 399–458.

G.L. Priest (1991), 'The Modern Expansion of Tort Liability: Its Sources, Its Effects and Its Reform', *Journal of Economic Perspectives*, vol. 5, pp. 31–50.

A. Rubinstein (1982), 'Perfect Equilibrium in a Bargaining Model', *Econometrica*, vol. 50, pp. 97–109.

S.W. Salant (1984), *Litigation of Settlement Demands Questioned by Bayesian Defendants*, Social Science Working Paper 516, California Institute of Technology.

S.W. Salant, and G. Rest (1982), *Litigation of Questioned Settlement Claims: A Bayesian Nash Equilibrium Approach*, Discussion Paper p6809, Rand Corporation.

W. Samuelson (1983), *Negotiation v. Litigation*, Discussion paper, School of Management, Boston University.

U. Schweizer (1989), 'Litigation and Settlement Under Two-sided Incomplete Information', *Review of Economic Studies*, vol. 56, pp. 163–78.

S. Shavell (1982), 'Suit, Settlement and Trial: A Theoretical Analysis Under Alternative Methods for the Allocation of Legal Costs, *Journal of Legal Studies*, vol. 11, pp. 55–81.

—— (1989), 'Sharing of Information Prior to Settlement or Litigation', *Rand Journal of Economics*, vol. 20, pp. 183–95.

III

TRADE AGREEMENTS AND INTELLECTUAL PROPERTY RIGHTS

III

TRADE AGREEMENTS AND INTELLECTUAL PROPERTY RIGHTS

6

REGIONAL TRADE AGREEMENTS

Building Blocks or Stumbling Blocks?

A. Jayagovind

6.1 THE ROLE OF REGIONAL TRADE AGREEMENTS

The tension between regionalism and centralism has been a perennial phenomenon in national systems. Most of the internal conflicts in different parts of the world can be attributed to the failure to strike a proper balance between these two forces. There is a general consensus that a successful polity requires regionalism to be accepted as a fact and that a durable federal structure be built by harmonizing regional aspirations.[1] Can this thesis be extrapolated into an international system? The answer is not certain. In the post-war world, the focus, at least in the initial decades, was on establishing global organizations to manage international relations, whether political or economic. The UNO as a political organization accorded a somewhat grudging recognition to regionalism in Article 51 of the Charter. The same was true in the context of the General Agreement on Tariffs and Trade (GATT). Article XXIV is the only provision in the entire GATT/World Trade Organization system recognizing some kind of role for regional economic organizations. But in practice, both in politics and economics, regional factors have proved to be too formidable for central organizations to cope with. This paper seeks to address

[1] The stability of the political systems in India, Canada, and the USA despite the diversities within these countries can be attributed to a great extent to successful federal model.

the role of regional trade agreements (RTAs) in the context of the multilateralism assiduously promoted by the GATT/WTO.

Economists like Jacob Viner[2] have been generally of the view that RTAs, to start with, inevitably lead to trade diversion from the efficient producers outside the trading blocs to the relatively inefficient producers within the blocs. But hopefully, there will be trade creation in course of time as a result of 'competitive trade liberalization.' In other words, trade liberalization within a trading bloc would spur liberalization worldwide in time. However, there is nothing inevitable about such a positive and optimistic development. Proliferation of trade blocs is as consistent with protectionism as with liberalization. As Sydney Dell put it:

> A policy of integration of Western Europe does not in itself imply adherence to the free trade idea as a general proposition. In other words, it does not by any means follow that, because one would like to see a larger and more powerful unit created out of many smaller states in Western Europe, one necessarily believes also in a policy of free trade for such a unit. Intense protectionism is just as much consistent with economic and political integration of Western Europe as intense anti-protectionism is.[3]

Article XXIV of the GATT is based on the premise that it is possible to harness regional aspirations to promote the ideal of multilateralism by channelling them through certain structured and functionally effective moulds. However, this idea, expected to sail smoothly, has encountered not a few difficulties.

6.2 RTAs AND THE GATT

In the course of negotiations for the proposed International Trade Organization, the USA was positively hostile towards any idea of preferential treatment. However, it was prepared to concede the benefits of a proper Customs Union (CU) or Free Trade Area (FTA). Clair Wilcox, the leader of the US delegation to the Havana Conference, distinguished a CU from a preferential system as follows:

> A Customs Union creates a wide trade area, removes obstacles to competition, makes possible a more economic allocation of resources, and thus operates to

[2] J. Viner, *Customs Union Issue* (New York: Carnegie Endowment for International Peace, 1950).

[3] S. Dell, *Trade Blocks and Common Markets* (Knoff, 1963), p. 82.

increase production and raise planes of living. A preferential system on the other hand retains internal barriers, obstructs economy in production, and restrains the growth of income and demand. A Customs Union is conducive to the expansion of trade on the basis of multilateralism and non-discrimination; a preferential system is not.[4]

The need, according to the USA, was to ensure that various RTAs would move towards free trade areas (FTAs) and eventually become Customs Union (CU). Article XXIV was drafted with a view to establishing a legal mechanism to facilitate and ensure such a transition. Article XXIV : 4 describes the rationale of RTAs as follows:

The contracting parties recognize the desirability of increasing the freedom of trade by the development, through voluntary agreements, of closer integration between the economies of the countries parties to such agreements. They also recognize that the purpose of a customs union or of a free-trade area should be to facilitate trade between the constituent territories and not to raise barriers to the trade of other contracting parties with such territories.

It is interesting to note that Article XXIV never uses the expression 'RTA', though all the delegates had RTAs in mind while approving this Article. They also believed that it would be possible to ensure that these CUs and FTAs would be building blocks for the proposed multilateral system. According to para. 7 of Article XXIV, the states that want to establish an FTA or a CU among themselves must notify their intention to the GATT and submit a plan and schedule delineating the roadmap towards that goal. If the Contracting Parties are not satisfied with the plan and schedule, or even with the steps thereof or period contemplated therein, they shall make recommendations for appropriate modifications to the plan and the schedule. 'The parties shall not maintain or put into force, as the case may be, such agreement if they are not prepared to modify it in accordance with the recommendations'.[5] Under Article XXIV: 10, the Contracting Parties may approve a plan and schedule by two-third majority, even if they do not comply with the requirements of Article XXIV, provided that they are convinced that such plan and schedule will eventually lead to a CU or FTA. In brief, Article XXIV is one of

[4] C. Wilcox, *A Charter for World Trade* (New York: Macmillan, 1949), pp. 70–1.

[5] The last sentence of Article XXIV : 7 (b)

the very few provisions of the GATT vesting specific powers in the Contracting Parties.

In order to approve a plan and schedule for CU or FTA, the Contracting Parties have to ensure that these three requirements are met:

1. Duties and other regulations of commerce imposed at the institution of the CU (or maintained at the institution of FTA) or interim agreement in respect of trade with states not parties to the proposed union shall not be on the whole higher or more restrictive than the general incidence of the duties and regulations of commerce that were applicable earlier.
2. Duties and other restrictive regulations of commerce are eliminated with respect to substantially all the trade between the constituent territories of the proposed union.
3. In case the parties wish to proceed towards the union via interim agreement, the time envisaged therein is reasonable.

Under the GATT, the parties to the proposed RTAs used to notify their plans and schedules to the GATT Council which in turn would appoint ad hoc working parties to go into these plans and schedules. During the GATT period, 27 working parties were so established to examine the roadmaps towards the proposed unions. Except in the case of a Canada-USA free trade agreement, the working parties could not submit the unanimous reports expected of them. Thanks to the principle of consensus which somehow became a binding convention in the GATT decision-making process, the Contracting Parties could not take any decision either approving or rejecting these proposals. There were always disagreements about the methodologies to be adopted for deciding 'general incidence of duties and other regulations' and also 'substantially all the trade' relations. There was no clear understanding as to 'reasonable length of time' within which the parties have to realize the goal of CU or FTA. In the absence of specific recommendations for modifications, thanks to the consensus requirement, the parties went ahead with their original plans.[6]

[6] T. Cottier, and M. Foltea, 'Constitutional Functions of the WTO and Regional Trade Agreements', in Bartels and Ortino (ed.), *Regional Trade Agreement and the WTO Legal System* (New York : Oxford University Press, 2006) pp. 42–74.

The provision for the approval by two-thirds majority under Article XXIV:10 is based on the assumption that the decisions under para.7 are taken on simple majority. Once the consensus principle was introduced willy-nilly into the GATT, para.10 did not make any sense.

The Understanding on the Interpretation of Article XXIV of the GATT, 1994, introduced as part of the WTO Agreement, made a serious effort to remove the above lacunae. Thus, as far as the concept of the general incidence of duties is concerned, it provides that the decision shall be based on an overall assessment of weighted average tariff rates and of customs duties collected. This assessment shall be made on the basis of import statistics for a previous representative period. It also provides that the reasonable length of time for the formation of the proposed union should normally be ten years and any further extension requires 'full explanation' by the parties concerned.

Both CUs and FTAs must substantially free all the trade between the constituent territories from duties and other restrictive regulations. Unless this is achieved, the RTA would be no more than a cleverly disguised protection of certain industries by member states. There has been much discussion regarding the appropriate criterion in this regard: whether it should be qualitative or quantitative or both. In the case of *Turkey: Restrictions on Imports of Textiles and Clothing Products*[7] both the Panel and the Appellate Body held that the criterion shall embody both qualitative and quantitative elements. The Understanding on the Interpretation of Article XXIV tackles this issue somewhat tangentially. In its preamble, it states that:

> Recognizing also that such contribution (to the expansion of world trade) is increased if the elimination between the constituent territories of duties and other restrictive regulations of commerce extends to *all trade* and diminished if any major sector of trade is excluded …

What is stated above is obvious, but its relevance to the interpretation of 'substantially all' the trade is not clear. Even within the European Economic Community, the only CU we have at present, trade in agricultural commodities is restricted.

Under the WTO dispensation, a permanent Committee on Regional Trade Agreements (CRTA) was established in 1996 by the

[7] WT/DS34/R 1999 and WT/DS34/AB/R.1999.

General Council, replacing the earlier practice of appointing ad hoc committees. This CRTA is expected to scrutinize all the proposals for RTAs with a view to determining their compatibility with Article XXIV; and for this purpose, it has developed elaborate procedures comprising oral and written presentations. At the end thereof, the CRTA would submit its reports to the Council for Trade in Goods and the Council in turn is expected to submit these reports with their own recommendations to the General Council.

There has been a virtual explosion of RTAs under the WTO. Nearly 162 RTAs have been notified to the Council for Trade in Goods since 1995 and the CRTA submitted the examination report in 118 cases. But none of these reports was accepted by the General Council in view of the consensus requirement. In the absence of any specific recommendations under Article XXIV: 7 (b), the parties concerned have gone ahead with their plans. In other words, the newly introduced 'Understanding' has not in any significant way contributed to the effective supervision of RTAs.[8]

It may be noted that both under Article XXV of the GATT and Article IX of the WTO, the decisions could be taken by voting. But the convention, developed informally under the GATT, was to take the decisions by consensus only, and the same continued under the WTO. There is some merit in this methodology given the decentralized international system; but it makes the enforcement difficult. In this context, para. 12 of the Understanding on Article XXIV may assume significance. It reads:

> The provisions of Articles XXII and XXIII of GATT, 1994, as elaborated and applied by Dispute Settlement Understanding may be invoked with respect to any matters arising from the application of those provisions of Article XXIV relating to CUs [Customs Unions], FTAs [Free Trade Areas], or interim agreements leading to the formation of a CU or FTA.

Can a member challenge a plan and schedule for an FTA under DSU on the grounds that it does not cover 'substantially all' the trade or on some other grounds discussed above? In other words, can a panel or appellate body under the DSU enter the scene and break the stalemate in which the General Council has got struck on account of consensus requirement? From a purely positivistic legal point of view, the answer

[8] T. Collier and M. Foltea (2006).

is 'yes'. The *India-Quantitative Restrictions* case is an example in this connection.[9] The Undertaking on the Balance of Payments Provisions of the GATT, 1994, contains exactly identical provisions regarding the dispute settlement. In this case, the Balance of Payments Committee did not make any recommendation for the removal of quantitative restrictions as per Article XVIII: Section B. When the matter was brought before the Panel and the Appellate Body, India developed an argument on the basis of institutional balance, that is, when a particular function is entrusted to a given organ of an institution, that function cannot be usurped by other organs of the institution. This contention was summarily rejected both at the panel and appellate level by giving the literal meaning to the expression that the DSU may be invoked with respect to *any matter* arising from the application of the relevant GATT provisions. If the same logic is pursued here as well, the expressions such as 'general incidence', 'substantially all', etc. can be subjected to adjudication. Whether such an approach is otherwise desirable from a policy point of view is another matter. Many WTO rulings on sensitive issues like subsidies have not been implemented, and the rulings on the RTAs are likely to meet the same fate.

Judicial function can be effective only within a defined legal framework backed by the general consensus in the society. Where the people are sharply divided over an issue, it is for the legislature to evolve consensus. Judges cannot impose their decisions as solutions. When the EEC was discussed in the GATT, it was pointed out that the EEC was more powerful than the GATT and it was the GATT which required support from the EEC and not vice versa. Under these circumstances, it is doubtful whether a panel can decide contentious issues involving RTAs which have formidable political dimensions as well.

6.3 PREFERENTIAL TRADE AGREEMENTS

In addition to the RTAs under Article XXIV, another type of free trade agreements have come up under the GATT Decision on Differential and More Favourable Treatment—Reciprocity and Fuller Participation of Developing Countries.[10] This Decision, popularly known as

[9] WT/DS/90/R 1999 and WT/DS 50/AB/R. 1999

[10] Decision of 28 November, 1979, L/4903 BISD 26th Supplement, 1978–79. pp. 203–4.

Enabling Clause was taken mainly to give effect to Generalised System of Preferences (GSPs) concluded under the auspices of the United Nations Conference on Trade and Development (UNCTAD) in 1971. Accordingly, para. 2 (a) of the Decision acts as a waiver from the obligation of most favoured nation treatment for those developed countries that wish to provide preferential tariff treatment to developing countries. In addition, under para. 2 (c), less-developed countries can enter into regional or global arrangements for the mutual reduction or elimination of tariffs. The agreements concluded hereunder are generally known as preferential trade agreements (PTAs).

PTAs are different from RTAs under Article XXIV in the sense that the former are not subject to the requirements such as substantial liberalization of the trade, etc. The only institutional obligation is to notify the Contracting Parties and consult with any affected contracting party. Less developed countries can come together to establish preferential trade areas without complying with the requirements of a CU or FTA and they are not subject to review as in the case of Article XXIV arrangements.

The PTAs have proliferated under the Enabling Clause. Among the developed countries, the European Community (EC) has made full use of these provisions. But while extending special treatment to developing countries, the EC sought to make tariff preferences contingent upon the pursuit of certain non-trade policies, such as protection of labour rights, environment and combating of drug trafficking. India successfully challenged tariff preferences conditional on combating drug production and trafficking.[11] The Panel, upholding India's contention, ruled 'the legal function of the Enabling Clause is to authorize derogation from Article 1:1 (of the GATT, 1994), a positive rule establishing obligations, so as to enable developed countries to provide GSP to developing countries. ... the Panel considers that this is a limited authorization of derogation in that the GSP has to be generalized, non-discriminatory and non-reciprocal.'[12]

Despite the above success, apparently the GSP has been of a limited utility to the developing countries. As the Sutherland Commission in

[11] European Communities: Conditions for Granting of Tariff Preferences to Developing Countries. WT/DS246/R 2003.

[12] Ibid., para. 7.38.

its report to the WTO put it: such GSP preferences were subject to few rules, with the selection of products, the level of preferences, and the choice of beneficiaries, left to the discretion of each developed country ... further it allowed for special and differential treatment with respect to rules on non-tariff measures.'[13]

6.4 SPAGHETTI BOWL

Jagdish Bhagwati coined the expression spaghetti bowl to refer to the overlapping of customs unions, regional and bilateral trade areas, preferences and innumerable trade deals between states. With the proliferation of so called RTAs, PTAs and bilateral trade deals, multilateralism has come under threat. In the words of the Sutherland Commission:

> Nearly five decades after the founding of the GATT, MFN is no longer the rule; it is almost an exception. Certainly much trade between the major economies is still conducted on MEN basis. However what has been termed the 'spaghetti bowl'. ... has almost reached the point where MFN treatment is exceptional treatment. Certainly, the term might now be better defined as LFN, Least Favoured Nation treatment.[14]

The Sutherland Commission used the term 'Preferential Trade Agreement' (PTAs) in a comprehensive sense to include agreements concluded under Article XXIV of the GATT, Article V of the GATS and the Enabling Clause. According to statistics, the total number of PTAs in operation was nearly 300 as of 2007.[15]

The type of preferential trade arrangements which have come up especially after 2000 defy the early categorizations. The USA has concluded number of bilateral FTAs with countries like Chile, Jordan, Egypt, Singapore, etc. India has concluded preferential trade agreements with Chile, Thailand, Singapore, and MERCOSUR. The expression Regional Trade Agreement has lost its significance in the context of these kinds of trade agreements. A close examination of the FTAs concluded by the USA would show that these agreements are no longer purely trade agreements: they contain provisions relating to investment, intellectual property rights and social clauses like labour rights, etc. In

[13] Sutherland Commission, *The Future of the WTO* (WTO, 2004), para. 70.

[14] Ibid., para. 60.

[15] Ibid., para.76.

relation to investments, they provide for national investment to the foreign investors even at the entry level.[16] With regard to IPRs, many of them contain provisions in addition to those provided in the TRIPs. All of them have been concluded under Article XXIV of the GATT.

In the present context of bilateral deals between the developed countries (like the USA) and the developing countries, the expression regional trade agreement has become an oxymoron: as in the case of the Holy Roman Empire which was neither holy nor Roman and far from being an Empire, the Regional Trade Agreements are neither regional, nor are they confined to trade and are far from agreements in real sense of the term. Many developing countries which have taken a public stance with regard to issues like investment, IPRs, social clauses, etc. have meekly accepted the stringent obligations under the bilateral deals and in this sense, the voluntariness which ought to be implicit in a true agreement is lacking in such arrangements.

Many reasons have been adduced for the proliferation of these kinds of agreements.[17] The apparent reason cited most of the time is the failure of successive WTO rounds to liberalize international trade. The positive explanation given in such a context is that those countries that want to forge deeper integration between their economies go for these kinds of deals. Many developing countries are driven by their anxiety to preserve their market share in the developed world. In many cases they are driven by security and political considerations. Invariably, the dominant motivation for developing countries is the hope of attracting foreign investment; and that is the main reason for including additional factors such as investment, IPRs, competition policies, etc. in the trade agreements.

The inclusion of TRIPs plus provisions in many bilateral FTAs concluded by the USA has given rise to some legal problems.[18] Such

[16] U.S. – Chile FTA, 2004. Article 12.2.:1: Each Party shall accord to investors of the other Party treatment no less favourable than that it accords to its own investors, in like circumstances, with respect to the establishment acquisition, expansion, management, conduct, operation and sale or other disposition of financial institutions and investments in financial institutions in its territory.

[17] WTO Secretariat, *The Changing Landscape of RTAs* (Geneva: WTO, 2003) paras 22–4.

[18] See Article 17(9)(4) of U.S.–Chile FTA 2004. U.S.–Morocco, U.S.–Singapore and U.S.–Bahrain are other examples.

provisions relate to limits on compulsory licensing, parallel imports, etc.[19] It is well known that the developing countries within the WTO are opposed to any extension of the TRIPs and the Doha Declaration on Public Health seeks to mitigate the impact of a strict interpretation of TRIPs provisions. The legal problem which some of the developing countries that are parties to these types of FTAs may face is that they may be compelled to extend the same treatment to other members of the WTO (especially to other developed countries) by virtue of Article 4 of the TRIPs that provides for the most-favoured nation treatment. The issue has not come up officially in any forum, but such a possibility cannot be ruled out. If a large number of developing countries become part of such arrangements, there may be an effort to push these issues to the WTO forum as well citing them as precedents.

A literal interpretation of Article 4 requires that whatever advantage a developing country has given in a IPR-related issue to another country must be immediately and unconditionally extended to all other members of the WTO. Article XXIV is a statutorily carved-out exception to the most-favoured nation treatment. But can Article XXIV of the GATT be an exception to Article 4 of the TRIPS? Para. 5 of Article XXIV speaks about reducing 'duties and other regulations of commerce' and one may argue that 'other regulations of commerce' should include all issues covered under various multilateral trade agreements, including the TRIPs. Though, historically speaking, Article XXIV was drafted only in the context of the GATT, there is no reason why the GATT, 1947, which is now re-designated as the GATT, 1994, should not take within its ambit all other issues such as IPRs which affect commerce. The virtue of such an interpretation is that the Committee on Regional Trade Agreements can scrutinize at least those issues like IPRs which fall within the WTO framework. And, if IPRs are covered by the expression 'other regulations of commerce', such concessions in IPRs need not be materialized.

[19] Mercurio, B., 'TRIPs-Plus Provisions in FTAs—Recent Trends', in Bartels and Ostino (eds), Regional Trade Agreements and the WTO Legal System (New York : Oxford University Press, 2006) pp. 215–37.

6.5 REGIONAL TRADE AGREEMENTS AS SIGNALLING THE FAILURE OF MULTILATERALISM

The expression 'regional trade agreement' is not used in the GATT, though the publicists have been consistently using this expression. There is enough evidence to indicate that the framers of the GATT had only RTAs in mind when they used expressions such as 'CUs' and 'FTAs'. In the present context where FTAs are concluded between the countries across continents, RTAs have lost their moorings altogether. They are neither regional, nor are they confined to trade and are far from agreements in the real sense of the term. Following the Sutherland Commission, we can use the expression 'preferential trade agreement' with a caveat that it should not be confused with the agreements under the Enabling Clause. Preferential agreements may be used as a generic expression to cover agreements under Article XXIV of the GATT, Article V of the GATT and the Enabling Clause.

Whether, from an economic point of view, RTAs have been building blocks or stumbling blocks is nearly impossible to answer. This is all the more so when they are treated as a second-best option in the absence of progress (even to the extent of regress) in multilateral trade liberalization. But from a legal and institutional point of view, the 'spaghetti bowl' phenomenon is threatening the edifice of multilateralism. The controlling mechanism (which was to lead them towards multilateral framework) as envisaged in para. 5, 6, and 7 of the GATT, 1994, has become moribund mainly on account of consensus rule. Going by precedents, it may be possible to bypass the consensus rule by resorting to a dispute settlement mechanism. But it may not be desirable to do so. Even with regard to purely economic issues such as subsidies, the WTO has failed to enforce the decisions given by the adjudicating bodies; and political dimensions associated with FTAs make the adjudication totally undesirable.

Para. 29 of the Doha Declaration mandates that negotiations must aim at clarifying and improving disciplines and procedures applicable to RTAs. The 'Understanding on Article XXIV' has clarified the issues to the extent possible. In the final analysis, RTAs may be considered as symptoms of the failure of multilateralism and therefore, efforts must be concentrated on reviving and strengthening multilateral agenda.

REFERENCES

Bartels and Ortino (ed.), *Regional Trade Agreements and the WTO Legal System*, New York: Oxford University Press, 2006.

Dell, Sydney, *Trade Blocs and Common Markets*, New York: Knoff, 1953.

Sutherland Commission, *The Future of the WTO*, WTO, 2004.

Viner, Jacob, *Customs Union Issue*, New York: Carnegic Endowment for International Peace, 1950.

Wilcox, Clair, *A Charter for World Trade*, New York: Macmillan, 1949.

WTO Secretariat, *The Changing Landscape of RTAs*, WTO, 2003.

7

INCREMENTAL INNOVATION AND PATENT PROTECTION FOR PHARMACEUTICAL PRODUCTS IN INDIA

A Law and Economics Analysis of the *Novartis* Case[1]

Thomas Eger, Petra Ebermann, and Padmanabha Ramanujam*

7.1 PATENTS IN INDIA

In 2005, patent protection in India underwent a dramatic change. Whereas the Patent Act of 1970 excluded product patents for food, medicinal drugs, and products of chemical processes from patentability, the 2005 Patents (Amendment) Act allows for product patents also in the pharmaceutical sector, with one important qualification: According to Section 3(d), new forms of existing pharmaceutical substances that do not result in significantly enhanced 'efficacy' or employ at least one new reactant are not patentable. This Act constitutes India's last step towards complete compliance with the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS Agreement).

There is some evidence that India is moving away from occupying the post of a mere producer of generics to an innovator developing new drugs and also improving existing drugs in order to make them more suitable for the specific conditions in less developed

[1] *Novartis AG* v. *Union of India and Others*, W.P. No. 24754 of 2006 and W.P. No. 24759 of 2006.

* We wish to thank Wolfgang Kerber, the participants of the senior scholar workshop in Hyderabad (February 2008) and an anonymous referee for valuable comments. The usual disclaimer applies.

countries.[2] This paper presents a study which intends to capture the widely discussed Indian patent policy for the pharmaceutical sector. We describe the legal framework of patent protection in India, which is to some extent determined by the TRIPS Agreement and other international agreements (Sec 7.2). Thereafter, we present the problem of incremental innovation with reference to the recent and controversially discussed Novartis case that centres on Section 3(d) of the 2005 Patents (Amendment) Act (Sec 7.3) and analyse it from a law and economics perspective (Sec 7.4).

7.2 LEGAL FRAMEWORK FOR PATENT PROTECTION

7.2.1 TRIPS as a Legal Framework

One of the problems with patents and international trade is that the principle of territory and independence of patents dictates that a patent can be granted in one country but have no protective effect in another. All intellectual property systems have an element of domestic regulation.[3] Patents are territorial rights granted by a sovereign government and valid only within the territory of the granting authority. As a consequence national patent systems focused for a long time on the optimal regulation of the national market. This lead to strong patent systems in countries with much innovative activity and weaker systems in countries with only a few potential innovators.[4] Trade in goods embodying intellectual property rights has substantially increased in the last decades and crossed national borders. This is one consequence of economic globalization which integrates national economies into one global economy.[5] With

[2] Sudip Chaudhuri, 6 August 2004, 'TRIPS, Indian Generic Companies and Accessibility of Medicines', Conference on Global Pharmaceuticals, Biotechnology and Health, Institute for Strategic Biotechnology, Health and Training, Mumbai. J.J. Thomas (2006), 'India and China in the Knowledge Economy: Rivals or Allies? Case Studies of Pharmaceuticals and Biotechnology', Paper presented at the Conference on China, India, and the International Economic Order, Singapore, June 23–4.

[3] B. Hoekmann and M. Kostecki (2001), *The WTO and Beyond*, 2nd edn, Oxford: Oxford University Press, p. 274.

[4] R.E. Falvey, F. Martinez, and G.V. Reed (2002), 'Trade and the Globalisation of Patent Rights', GEP Research Paper No 02/21, available at http://ssrn.com/abstract=413263 (last accessed on 10 July 2009).

[5] P. Van den Bossche (2007), *The Law and Policy of the World Trade Organisation*, Cambridge: Cambridge University Press.

globalization and the growing importance of international trade and international markets the demand for having the same or similar protection in foreign markets as in the national market has increased.[6] The globalization of the international economy generated a desire for the harmonization of intellectual property rights, including patents, on an international level. There have been attempts to have multilateral cooperation in the field of intellectual property protection, for example the Berne Convention, or the Paris Convention for the Protection of Industrial Property administered by the World Intellectual Property Organization (WIPO), but they proved to be insufficient to achieve greater harmonization.

The differences in intellectual property systems that prevailed at the national level with regard to the granting and enforcing of intellectual property rights acted as non-tariff barriers to trade. This led to the formation of TRIPS, which established some degree of harmonization in the area of intellectual property rights at the international level.[7] The TRIPS Agreement was negotiated during the Uruguay round under the auspices of the World Trade Organisation (WTO) and establishes inter alia minimum standards for patent protection. These standards apply to availability, scope, and use of intellectual property rights, including enforcement and process requirements. General principles of the General Agreement on Tariffs in Trade (GATT) such as the most-favoured nation clause, national treatment and non-discrimination are part of TRIPS as well (Article 3 of TRIPS). The TRIPS Agreement incorporates rules and obligations from existing treaties covering intellectual property and in addition introduces a dispute settlement mechanism.[8] Under this mechanism a country whose rights have been infringed under one agreement may suspend obligations under a different agreement, thus providing for retaliation across agreements and sectors. It is worth noting that the TRIPS Agreement is unique in the WTO context. The agreements administered by the WTO have so far only been concerned with the issue of freeing trade by providing

[6] Falvey, *et al.* (2002), p. 15.

[7] Ibid., p. 2.

[8] A. Kamperman Sanders (2007), 'Intellectual Property Law and Policy and Economic Development with Special Reference to China', in: T. Eger, M. Faure and Z. Naigen (eds), *Economic Analysis of Law in China*, Cheltenham, pp. 239–71.

for the exclusion or penalization of the use of certain policies that are presumed to affect trade negatively (negative integration). TRIPS, on the other hand, imposes obligations on national governments to adopt a set of substantial rules in a field that is commonly considered to be subject of domestic regulation, having a direct effect on national regulatory and legal regimes (positive integration).[9] The Member States are free only in as far as they can determine how the specified minimum standards are implemented (Article 1 of TRIPS).

With regard to patents, Section 5 Part II (Article 27–34 of TRIPS) regulates the availability, scope, and use of patents. Article 27 (1) states that patents shall be available for any invention, whether products or processes, in all fields of technology, provided that they are new, involve an inventive step, and are capable of industrial application. The terms 'non-obviousness' and 'utility' may be deemed to be synonymous with 'inventive step' and 'capable of industrial application' (Article 27 FN 5 TRIPS). Some countries made use of this provision and speak of non-obviousness and utility in their patent laws.

Unlike most other patent laws, the TRIPS Agreement does not define what is considered an invention. Rather, it relies on the eligibility criteria of novelty, inventiveness, and capability of industrial application.

Novelty requires that a claimed subject matter is not part of the state of the art at the time of filing the patent application. The state of the art encompasses everything that is available to the public. Differences in national laws can be found in regard to the scope of the prior art. While most countries include in the relevant prior knowledge everything that has been made available to the public regardless of the form and place of disclosure, the United States determines prior knowledge of unpublished information on a domestic scale. Unpublished knowledge forms part of the state of the art only if it originates from within the United States. A patentable invention must also show an inventive step or non-obviousness of the alleged invention. This requirement is met when a person skilled in the art under consideration does not regard the invention obvious and excludes all those inventions that are merely the next logical step or natural consequence of a former invention. Finally, an invention must be capable of industrial application or

[9] Hoeckmann and Kostecki (2001), p. 274.

useful. Industrial application includes trade, industry and agriculture. Usefulness requires the invention to have some current beneficial use to the public.

The minimum term of protection that needs to be provided for by domestic patent laws is 20 years from the date of filing. Exceptions to the general obligation to provide patent protection can be found in Article 27 (2) and (3) of TRIPS. Members can exclude inventions from patentability in case the *ordre public* or morality considerations make such exclusions necessary. Furthermore, diagnostic, therapeutic, and surgical methods for the treatment of humans or animals as well as plants and animals other than micro-organisms, and essentially biological processes for the production of plants or animals other than non-biological and microbiological processes may be excluded from patentability.

Globalization in general, and the TRIPS Agreement in particular, especially the provisions on obligatory patent protection have been subject to a heated North-South debate. While the developed countries at large demanded strong patent protection by the WTO, most developing and underdeveloped countries were concerned that stronger patent protection would be detrimental to their economic development and welfare. Another point of discussion—the adequate protection of incremental innovations—is discussed in detail in the following sections.

7.2.2 Patent Law in India

Although the objective underlying the Indian Patent Act of 1970 was to encourage inventors to invest in their creative abilities knowing that their inventions would be protected by law,[10] product patents for food, medicinal drugs, and products of chemical processes were excluded from patentability under Section 5 of the Patents Act of 1970 in order to facilitate access to essential medicine.

Prior to the Patent Act of 1970 the market for pharmaceuticals was dominated by Multi-National Companies (MNCs). By using the

[10] In the case of *Bishwanath Prasad Radhey Shyam* v. *Hindustan Metal Industries*, (1979) 2 SCC 511 at 517, the Supreme Court of India ruled that, 'the object of patent law is to encourage scientific research, new technology and industrial progress.'

existing patent laws, the MNCs prevented home-grown companies from producing new drugs. This led to an increase in drug prices due to a lack of competition. This situation was completely transformed with the introduction of the Patent Act of 1970. The Patent Act introduced a provision which removed product patent protection and thus allowed for process patents only.[11] This change in patent law brought about significant changes in the Indian pharmaceutical industry as it allowed Indian pharmaceutical companies to 'reverse engineer' the manufacturing process of existing drugs and thus develop generic drugs. As a consequence of this legislation, large scale production of bulk drugs by the native pharmaceutical industry developed particularly in the 1980s. This is considered to be the most important achievement of the Indian pharmaceutical industry.[12] Further rapid growth and intensification of the pharmaceutical sector took place in the 1990s, which is accordingly considered to be the golden era of the Indian pharmaceutical industry. Today India supplies 8 per cent of the total global output (in volume) of drugs and 22 per cent of the world output of generic drugs.[13] There has been a significant increase in exports since the mid 1990s, with the US taking up the role of the most important export partner.[14]

Like all other member countries, India entered into trade obligations and agreed to further comply with the provision of the TRIPS Agreement when it joined the WTO in 1995. Accordingly, the Patent (Amendment) Act of 2005 deleted Section 5 of the Patent Act of 1970 and allowed the grant of product patents for food, medicinal drugs, and products of chemical processes in consonance with Article 27 of TRIPS.

But the Patent Amendment Act of 2005 also introduced Section 3(d), which is unique in patent law and is the subject of a detailed discussion in this paper. Section 3(d) of the Act deems

[11] J. Watal (2008), 'Patents', in K. Basu (ed.), *The Oxford Companion to Economics in India*, revised edition, New Delhi: Oxford University Press, pp. 400–3.

[12] Thakore (1989); Raizada (2002).

[13] Thomas (2006), p. 14.

[14] S. Chaudhuri (2007), 'Is Product Patent Protection Necessary in Developing Countries for Innovation? R&D by Indian Pharmaceutical Companies after TRIPS', Working Paper No. 614/September 2007, Indian Institute of Management, Calcutta, p. 10.

'the mere discovery of a new form of a known substance which does not result in the enhancement of the known efficacy of that substance, or the mere discovery of any new property or new use for a known substance or, of the mere use of a known process, machine, or apparatus unless such known process results in a new product or employs at least one new reactant,' to be not patentable.

This provisions acts as a barrier to patenting products that differ only slightly from existing products, and is especially relevant with regard to pharmaceutical patents. In the recent Novartis case the Madras High Court order rejected the Novartis challenge of Section 3(d) thus reopening the debate as to whether disallowing patents on incremental innovation will discourage firms from doing research that builds upon existing inventions.

7.3 INCREMENTAL INNOVATION AND THE *NOVARTIS* CASE

7.3.1 What is Incremental Innovation?

An innovation is generally defined as an idea, practice, or material artefact perceived to be new by the relevant unit of adoption.[15] The degree of novelty that is captured by the notion of radical innovations in scientific literature also forms an essential requirement for granting patents. The literature on management science studies and classifies innovations under two categories namely (i) radical innovations and (ii) incremental innovations. Radical innovations are defined as innovations with fundamental changes that represent revolutionary changes in technology and show obvious departures from existing practice.[16] In contrast, incremental innovations are minor improvements or simple adjustments on the existing technology.[17]

The pharmaceutical industry provides an excellent setting to test determinants of incremental and radical innovation. Radical innovation refers to the identification of new chemical entities and

[15] N. Zaltmang, *et al.* (1973) *Innovations and Organizations*, New York: John Wiley & Sons.

[16] T.D. Duschesneau, S. Cohn, and J. Dutton (1979), 'A Study of Innovation in Manufacturing, Determination, Processes and Methodological Issues', vol. I. Social Science Research Institute, University of Maine. J.E. Ettlie (1983), 'Organizational Policy and Innovation among Suppliers to the Food Processing Sector', 26, *Academic Management Journal*, pp. 27–44.

[17] R.D. Dewar and J.E. Dutton (1986), 'The Adoption of Radical and Incremental Innovations: An Empirical Analysis', *Management Science*, vol. 32, pp. 1422–33.

their development into potentially useful pharmaceutical drugs. Incremental innovation, on the other hand, works with already known chemical compounds that are merely altered or employed in a different use.[18]

Drug enhancements such as new dosage forms may appear at first sight to be unimportant or even trivial but they are important avenues of learning for firms. Incremental progress gives rise to families or classes of related drugs. Although several agents within a class may have the same general action, they often differ significantly in specific actions, side effects, and suitability for individual patients.[19] Consequently, incremental innovation takes many forms, including improved safety and effectiveness, fewer side effects, new formulations allowing greater ease of use and improved compliance, new indications, and new versions of the medicine developed for specific groups of patients (such as children).[20] It can also take the form of greater product stability during storage and transport which can be especially important in tropical climates like some regions of India.[21] Thus in the pharmaceutical sector incremental innovation connotes the continuous improvement of medicines, which also requires large-scale research and development, including clinical trials, along with approval from regulators before the new product can be offered to patients. It therefore becomes important to afford protection to such innovations.

7.3.2 The Importance of Incremental Innovation

For many years radical innovation had been the primary goal of research for firms in many areas of science and technology. However, breakthrough innovations are important but rare in medical research.

[18] K. Cool (1985), *Strategic Group Formation and Strategic Group Shifts: A Longitudinal Analysis of the U.S. Pindustry, 1963–1982*, Unpublished Doctoral Dissertation, Purdue University, West Lafayette, IN.

[19] R.A. Levy (1990), 'Pharmaceutical Research: Therapeutic and Economic Value of Incremental Innovations', Unpublished Dissertation, National Pharmaceutical Council, Reston, VA. C.M. Banbury, W. Mitchell (1995), 'The Effect of Introducing Important Incremental Innovations on Market Share and Business Survival', *Strategic Management Journal*, vol. 16, pp. 161–82.

[20] Pharma Press Release, 2007.

[21] Ibid.

Most medical advances—like in all other technological fields—happen by 'incremental innovation' that is, innovation that builds on previous inventions. In the last 20 years, a number of noticeable changes have taken place in the type of research undertaken in all industries, and pharmaceutical industries in particular. These changes were motivated by the realization that due to long-run time horizons, high failure rates, and a low probability of returns the possibility of discovering new drugs was decreasing.[22] As a result, the focus of research shifted and concentrated on the discovery of new uses of known substances.[23] The problem that confronted researchers working in the area of incremental innovation was that traditional patent law refused to recognize the discovery of new advantages of an old product as being novel. Lionel Bently and Brad Sherman conceptualize this problem in the following example. Assume that someone discovered and patented aspirin as a drug useful for curing headaches. Later someone else found out that the consumption of aspirin also thinned the blood and was thus useful in preventing blood clots. The second use would not be patentable due to the fact that aspirin is already patented as a drug for curing headaches.[24] The reason for this is that traditional patent law in many countries treated a claim to a 'product for a particular use' as a claim to the product per se; consequently the product would lack novelty even if it had previously been employed in a different use.[25]

One of the prominent trends in recent years is the way in which this principle has slowly been diluted in many countries, at least in areas of therapeutic research, highlighting the economic and strategic importance embodied in new use patents. The standard reasoning used by the supporters of new use patents is that the human and

[22] S. Min, M.U. Kalwani, and W.T. Robinson (2006), 'Market Pioneer and Early Follower Survival Risks: A Contingency Analysis of Really New Versus Incrementally New Product-Markets', *Journal of Marketing*, vol. 70, pp. 15–33. S. Bhaskaran (2006), 'Incremental Innovation and Business Performance: Small and Medium-Size Food Enterprises in a Concentrated Industry Environment', *Journal of Small Business Management*, vol. 44, pp. 64–80.

[23] Cool (1985); Levy (1990); and Banbury and Mitchell (1995).

[24] L. Bently and B. Sherman (2001), *Intellectual Property Law*, New York: Oxford University Press, p. 8.

[25] This applies only if the product was used in the same technical art. Some jurisdictions also allow for second-use innovations in the same art, for example, the Swiss-claim-formula applied by the European Patent Office.

economic investment incurred by the industry to discover a new use is comparable to the investment necessary to the development of the product itself.[26] It will be interesting to investigate the international position on improvement and on new use patents which provide protection for such incremental innovations.

7.3.3 The International Position on Incremental Innovation

In the area of incremental innovation, we find different forms of innovations such as the improvement of a known substance, the combination of two inventions, or the discovery of a new use of a known substance. Generally speaking, incremental innovation is, under most patent laws, eligible for patent protection, provided that the requirements of novelty, inventiveness, and industrial applicability are met. The most difficult and decisive hurdle is meeting the inventive step (or non-obviousness) requirement.

7.3.3.1 Improvements

As for improving known and patented subject matter the difficulties lie in the assessment of the degree of inventiveness required to constitute a new patentable product. Ideally patent protection increases with the amount of innovation embodied in an invention.[27] The assessment of inventiveness is thus subject to Patent Office policy, guidelines, and caselaw. National laws take different approaches as to how to assess inventiveness.

United States: United States patent law is regulated in the Patent Code 35 USC Section 101 of the US Patent Act deems any new and useful improvement of a new and useful process, machine, manufacture, or composition of matter patentable if it complies with the criteria that determine patentability. Section 103 incorporates the requirement for an invention to be non-obvious in order to be patentable and states that a patent cannot be obtained 'if the differences between the subject matter sought to be patented and the prior art are such that

[26] Ph. W. Grubb (2004), *Patents for Chemicals, Pharmaceuticals and Biotechnology: Fundamentals of Global Law, Practice and Strategy*, 4th edn, Oxford: Oxford University Press.

[27] V. Denicolò (2008), 'Economic Theories of the Nonobviousness Requirement for Patentability: A Survey', *Lewis & Clark Law Review*, vol. 12, pp. 443–59.

the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains.' The content of this requirement is determined by case law. The leading case is *Graham* v. *John Deere*[28] which determines a test for non-obviousness. The test consists of four factual inquiries. The so called 'Graham Factors' are the scope and content of the prior art, differences between the claimed invention and the prior art, the assessment of whether the differences would have been obvious to a person skilled in the art, and finally, secondary considerations such as commercial success, failure of others, or a long felt need for improvement. The standard set by *Graham* v. *John Deere* was subsequently lowered. Instead of relying on the Graham test alone, courts applied the teaching-suggestion-motivation (TSM) test, finding obviousness only if some suggestion or motivation in the prior art suggested the invention (for example, *Winner Int'l Royalty Corp.* v. *Wang*).[29] Recently, however, a return to a tougher standard can be observed. In 2007 the Supreme Court reaffirmed the Graham test in the judgment in *KSR International Co* v. *Teleflex Inc.*,[30] holding that while *Graham* and TSM are not necessarily inconsistent, it is the Graham test and not the TSM test that is the appropriate basis to determine obviousness. *KSR* concerned a mechanical patent; it remains to be seen whether the reasoning will be applied to the pharmaceutical arts, but there appears to be a tendency that courts will do so. In regard to pharmaceutical patenting the decision in *Bayer Schering AG* v. *Barr Laboratories, Inc.*[31] is worth mentioning. The applicant Bayer held a patent on the formulation for an oral contraceptive consisting of a combination of known agents with an improved coating. The court referred to the reaffirmation of the *Graham* inquiry by the Supreme Court and invalidated the patent on the grounds that the combination of known substances was obvious to try for a person skilled in the art.

Pharmaceutical innovation often occurs by incremental modification of prior art compounds. The development of a pharmaceutical

[28] 383 US 1 (1966).
[29] 202 F. 3d. 1340, 1348 (Fed. Cir., 2000).
[30] 127 S. Ct. 1727.
[31] 2008 WL 628592 (D.N.J.).

compound with a structure and utility similar to the prior art often establishes a case for obviousness which must be rejected by the patentee, thereby shifting the burden of proof.[32] However, in such a scenario it is not per se impossible to receive a patent. Rather *prima facie* obviousness can be rebutted by showing that the new pharmaceutical product that combines known compounds, or builds upon existing inventions, and differs only slightly in its chemical structure has unexpected properties, or shows a considerable difference in the degree of the same property, and that this outcome was not suggested by the prior art or could not be anticipated or reasonably expected.[33] Here the US approach differs from the Indian patent law that requires for the provision of at least one new reactant.

Europe: European patent law is laid down in the European Patent Convention (EPC). Article 52 (1) EPC states that a patentable invention must involve an inventive step. According to Article 56 'an invention is considered to involve an inventive step if the invention, having regard to the state of the art, is not obvious to a person skilled in the art'.

The European approach to determine inventiveness differs only marginally from the US approach to non-obviousness. The differences relate to the prior art to be included when determining non-obviousness and to the question as to how to assess invalidating prior art. European patent law excludes previously filed but not yet issued patent applications from the prior art.[34] The European Patent Office (EPO) applies the problem-and-solution approach in assessing inventiveness (see EPO Guidelines). This approach addresses three issues: the determination of the closest prior art and the differences between the invention and the closest prior art, the formulation of the technical problem to be solved, and finally the question whether the claimed invention was obvious to a person skilled in the art starting from the closest prior art and the technical problem. The question is not whether the respective person could have arrived

[32] A.V. Trask (2008), 'Obvious To Try: A Proper Patentability Standard in the Pharmaceutical Arts?' *Fordham L. Rev.*, vol. 76, p. 2625.

[33] P. S. Ducor (1998), 'Patenting the Recombinant Products of Biotechnology and Other Molecules', The Hague: Kluwer Law International, p. 30.

[34] R.P. Merges and J.F. Duffy (2007), *Patent Law and Policy*, 4th edn, Newark, Law International, NJ: LexisNexis, p. 773.

at the solution but rather whether s/he would have done so (could-would approach).

Improving a known substance constitutes a valid patent claim if the requirements for inventiveness are met. The new product must not form part of the state of the art. It must differ significantly from the old product in a technical sense, for example, by a new formulation, dosage, or synergistic combination.[35] Furthermore the new improved product needs to be non-obvious in that sense the a person skilled in the art would not have made the improvement because s/he did not anticipate the positive effect.[36] The same applies to inventions that consist of a combination of known compounds.

Similar to US patent law, an improved product is patentable under the EPC if the research path or the outcome were not anticipated by the prior art. Consequently, it is not necessary to provide an increased efficacy but rather unexpected and sufficiently different properties.

7.3.3.2 New use

With regard to new (or second) use patents the following observations can be made. The identification of a medical indication of a known medical product cannot lead to patent claim for the product as such under the general principles of patent law. The question is whether the claim can encompass the use itself. The TRIPS agreement does not provide guidelines to Member States regarding patent protection for new uses of known products. Instead it leaves room for interpretation. Comparative law takes different approaches to this question.

United States: In the United States the patenting of new use inventions is possible provided that the purpose of the use is novel and non-obvious. In the assessment of obviousness, prior art can invalidate the patent only if the technology originated in an analogous art.[37] Consequently, new uses can be novel and non-obvious if the invention applies knowledge from a different technical discipline. In the case of method inventions the key consideration to determine patentability is

[35] R. Schulte (2001), *Patentgesetz*, 6th edition, Cologne: Heymanns, 2007, S. 1 Rn 190.

[36] Ibid., S. 1 Rn 191.

[37] J.R. Thomas (2005), *Pharmaceutical Patent Law*, Washington DC: BWA Books, p. 145.

whether the invention could be anticipated by other methods. Patents on uses are confined to a particular 'method-of-use' which does not encompass protection of the product as such (UNCTAD Resource Book on TRIPS and Development).[38]

Europe: Under Article 52 of the EPC patents are granted for new inventions which involve an inventive step and are capable of industrial application. Article 54 (5) of the EPC provides for the patentability of a new purpose of a known product. For instance, the identification of the first medical indication of a known product that did not have a pharmaceutical use before would be eligible for a patent claim for the product itself. Secondary uses of known products are not patentable when the claim concerns directions for the use of a product for the treatment of an illness. Such claims contravenes the directions laid down in Article 52 (4) EPO which deems methods of treatment unpatentable. However, claims directed to the use of a product for the manufacture of a medicament for a new specific therapeutic use can be subject to patent protection (so called Swiss-claims, EISAI OJ EPO 1985, 64).

Although the EPC covers European patents, some differences in the national laws of the Member States can be observed. New use patents appear to be granted reluctantly in the Member States.

United Kingdom: Courts in the United Kingdom are reluctant to approve new or secondary use patents. Although the principal patentability of secondary use claims is recognized since the decision in *Wyeth & Schering's Applications*[39] it has been difficult to maintain the validity of a secondary use patent before the courts. The secondary use of a medical compound may be patentable if the use is sufficiently novel as compared to the known use.[40] However, there is a tendency that such a new use will be considered a 'method of treatment' which would be excluded from patentability under Article 52 (4) EPC

[38] UNCTAD-ICTSD (2004), 'Resource Book on TRIPS and Development', Cambridge: Cambridge University Press, p. 356.

[39] (1985) RPC 545.

[40] M. Burdon and K. Sloper (2003), 'The Art of Using Secondary Patents to Improve Protection', *International Journal of Medical Marketing*, 3(3), pp. 226–38.

(*Bristol Myers Squibb* v. *Baker Norton*).[41] Only few secondary use patents have been approved in recent years, for example *SmithKline Beecham* v. *Generics*[42] and *SmithKline Beecham PLC* v. *Apotex Europe Ltd & Others*.[43]

Germany: German patent law allows for product patents on products for the first medical use. The second medical use can only be covered by a use-claim.[44] In this context it is worth noting that the Federal Patent Court in Germany invalidated a patent on secondary use in a recent decision (Bundespatentgericht 3 Ni 36/05 (EU)). The patent had previously been confirmed by the European Patent Office. The Federal Patent Court, however, held that the criteria for novelty set by the European Patent Office cannot bind the German courts. In case this decision is upheld in the higher instances it will be even more difficult to maintain the validity of secondary European Patents in Germany than, for example, in the United Kingdom.

It is difficult to define the degree of inventiveness required in order to patent incremental innovation. The difficulties lie mainly in the fact that much of the assessment is determined by case law and the results of some cases appear to be contradictory, since the individual circumstances matter and are taken into consideration by the courts. Nevertheless the overall trend is to allow patents on incremental innovation in pharmaceuticals but also demand some sort of unexpected and non-obvious result, be it new properties or an enhanced degree of the same property. It appears that Section 3(d) of the Indian Patent Act could suggest a similar approach. However, the Madras court interprets this section in a way that sets a tougher standard with regard to patenting incremental innovation than US and European patent law. This can be concluded from reviewing the recent decision in *Novartis* v. *Union of India*.

7.3.4 The Novartis Case (*Novartis* v. *Union of India)*

Novartis developed the innovative anti-cancer medicine Glivec, which is deployed for the treatment of chronic myeloid leukemia.

[41] (2000) ENPR 230.

[42] (2002) 25 (1) IPD 25005.

[43] (2003) 26 IPD 26020.

[44] G. Benkard (2006), *Patentgesetz*, Munich: Beck S., 5 Rn 49.

The base compound of Glivec is Glivec (imatinib), which is patented in 40 countries worldwide. Glivec (imatinib) is merely a research substance with no pharmaceutical form and cannot be taken by patients. It was only the first step in the development process of Glivec. The improvement embodied in Glivec is the transformation of the base compound into a pharmaceutically useful drug with a greater bioavailability than the base compound which can more efficiently be employed for the treatment of patients.

Although Novartis did not file a patent application for Glivec (imatinib) at the time of development in India since Indian patent law did not grant product patents for pharmaceutical products at that time, such a claim was made in January 2006 after the introduction of product patents in Indian patent law. However, the patent was denied under Section 3(d) of the Indian Patent (Amendment) Act 2005 because the substance was considered not to possess the required 'improved efficacy.' It was the contention of the Indian Patent Office that Glivec did not satisfy the requirements of novelty and inventiveness. Novartis challenged the denial of the patent for Glivec before the Madras High Court on the grounds that the decision lacked legal or factual basis and justification.

The learned senior counsels appearing for Novartis raised two arguments. Firstly, they challenged the constitutionality of Section 3(d) and claimed an infringement of the TRIPS obligations as well as the right to equality as laid down in Article 14 of the Indian Constitution. The second argument refers to the definition of an invention. It was asserted that the amended section is bad in law because the mere discovery of a new property is not treated as an invention unless the discovery of the new form of the known substance results in the enhancement of the known efficacy of that substance. Furthermore, it was argued by the appellants that the amended section, in the absence of any guidelines to determine the enhancement of the known efficacy of the substance on which the discovery relies, vests unguided discretion in the hands of the statutory authority and as a result is bad in law.

The Madras High Court decreed that the said Section 3(d) is not in violation of Article 14 of the Indian Constitution and found that it did not have jurisdiction to decide on the TRIPS issue.[45] Regarding the

[45] This argument is only mentioned for reason of completeness and will not be

interpretation of Section 3(d) the Court found that 'if the discovery of a new form of a known substance must be treated as an invention, then the Patent applicant should show that the substance so discovered has a better therapeutic effect'.[46] The Madras High Court concluded that the amended section sought to prevent 'evergreening' of patents,[47] provide easy access to the citizens of this country to life-saving drugs, and discharge the constitutional obligation of providing good health care to its citizens.

This decision of the Madras High Court removes many innovations that classify as incremental improvements from the scope of patent protection and supports the view that patents on incremental innovations are likely to be used by pharmaceutical firms to 'evergreen' their products by extending their monopoly period products. As a result, under Section 3(d) of the 2005 Patents (Amendment) Act, new uses of known substances cannot qualify as patentable, on the grounds that they do not fulfil the 'inventive step' requirement.

There are examples of other patent applications that were rejected under Section 3(d). In *M/s. Astra Aktiebolag*,[48] for example, the patent controller held that the claimed invention was not patentable under Section 3(d) as the benefit claimed by the applicant did not accrue to the user 'in terms of therapeutic quality of the product but to the manufacturer only in terms of consistency in the production of formulation.' In another patent application[49] the patent was refused

---

analysed in further detail.

[46] *Novartis* v. *Union of India*, 2007(4) MLJ 1153, 1290.

[47] 'Evergreening' is a term used to refer loosely to inappropriate extensions in the period of patent exclusivity for a pharmaceutical product.[...] Typically, it denotes a set of practices by patentees, wherein largely trivial or insignificant changes are made to a patented pharmaceutical product and then a secondary patent is applied for. If such a patent is granted and if a generic product on the market is modified to include the features mentioned in the secondary patent, the monopoly of the patentee is extended beyond the period of the first patent.' (S. Basheer (2005), 'Limiting the Patentability of Pharmaceutical Inventions and Micro-Organisms: A TRIPS Compatibility Review', Mimeo, 39). With regard to Section 3 (d) Indian Patent Act see ibid., 42, and S. Basheer, and T.P. Reddy (2008), 'The "Efficacy" of Indian Patent Law: Ironing out the Creases in Section 3(d)', in *Scripted*, 5(2), pp. 232–66, available at http://ssrn.com/abstract=1086254 (last accessed on 25 May 2009).

[48] Patent Application No. 1354/Del/1998, decision dated 12 June 2007.

[49] Patent application No. 841/Del/1996, decision on 30 August 2007.

under the provisions of Section 3(d); it was argued by the patent controller that although the claim possessed some improvements in the physical property of the drug, these improvements did not result in any new therapeutic effect.

In the aftermath of the decision of the Madras High Court the impact of Section 3(d) for the development of the pharmaceutical industry and on public health are still unclear. After the changes of the Patent Amendment Act became effective, the Indian pharmaceutical market experienced an increase in drug prices as well as a decrease in the number of pharmaceutical companies. A detailed survey of this development has yet to be undertaken and it is unclear whether the developments can be attributed to changes in patent policy. Some observers argue that the existence of Section 3(d) of the Indian patent law gives a wrong signal to the domestic as well as the international investing community and as a result will act as a barrier for R&D and innovation in the pharmaceutical sector in India.[50] There are others who argue that Section 3(d) forces firms to focus efforts on innovation and hardcore research instead of trifling with known substances and thereby sought to prevent 'evergreening' of patents.[51] The Madras High Court decision in the *Novartis* case has already taken the later position, but it is far from clear that this is the 'good position' in law. The way Section 3(d) is interpreted by the courts and patent office clearly demonstrates that the importance of incremental innovation in the pharmaceutical industry is being overlooked by defining 'efficacy' as 'therapeutic efficacy'.[52] Furthermore the courts have also failed in laying down proper guidelines as to how to determine the enhanced efficacy.

[50] See Ajit Dangi (Director General, Organization of Pharmaceutical Producers in India) available at http://www.rediff.com/ money /2007/aug/24deb.htm (last accessed 18 January 2008).

[51] See D.G. Shah (Vision Consulting Group) available at http://www.rediff.com/ money /2007/aug/24deb.htm (last accessed 18 January 2008).

[52] S. Basheer and T.P. Reddy (2008) and S. Basheer (2008), 'The "Glivec" Patent Saga: A 3-D Perspective on Indian Patent Policy and TRIPS Compliance. Paper presented at the Indo-German Conference on Intellectual Property Law', Freiburg and Munich, May 12–15, 2008, p. 14, available at http://www.atrip.org/upload/files/essays/Shamnad%20Basheer%20Glivec%20Patent%20Saga.doc (last accessed on 25 May 2009).

7.4 A LAW AND ECONOMICS PERSPECTIVE

From an economic perspective there are some good reasons in favour of the grant of patents, especially in the pharmaceutical industry. A patent grants an inventor a legal monopoly in the exploitation of his invention for a limited period of time. This serves the purpose of encouraging innovation and making them available for the public by giving the innovator an instrument to recoup his investment in the innovative activity. The subject matter protected by patents is basically information and as such not rivalled and easy to copy. Without a system that protects inventions there would be inadequate incentive to either innovate or to disclose innovations—to the detriment of the general public. Intellectual property rights are allocated to inventors in order to prevent expropriation and to provide for an environment in which bargaining and market exchange can take place.[53] Innovative work requires the input of large resources. R&D in this area does not only take long but is also extremely costly. Patent protection enables innovators to charge a price for innovative products above the marginal cost and thereby to recoup R&D expenses. The protection of inventions is of great importance, especially in the pharmaceutical sector due to the vast efforts and high risks involved in the process of developing a new and useful product. The results of Mansfield's detailed study on how patents affect industries suggested that patent protection was essential for the development of 30 per cent or more of the inventions in only two industries, namely the pharmaceutical and the chemical industry.[54]

However, to achieve the social goal of patent protection, and promote dynamic efficiency by inducing firms to engage in socially beneficial R&D activities, society has to pay a price. The social costs of patent protection consist of the static welfare losses due to the mark up on the marginal cost of producing the product that embodies the invention and, depending on the specific circumstances, the waste of resources resulting from patent races and of the increased cost of secondary innovation which could otherwise build on the original innovation.

[53] A. Kamperman Sanders (2007), p. 244.

[54] E. Mansfield (1986), 'Patents and Innovation: An Empirical Study', *Management Sciene*, 32 (2), p. 174.

Innovative activity is to a large extent a cumulative process.[55] Present innovations depend on past innovations. How should incremental innovations be treated by patent law from an efficiency point of view? A trade-off has to be solved between inducing investment in the original invention and inducing investments in improvements.

Let us denote the per-period value of some original (first) innovation to all potential final users v_1 and the required R&D investment c_1. Assuming that the innovation is a new product, such as a new drug, and the market for this new good is characterized by the demand function in Figure 7.1. For the sake of simplicity, without loss of generality, we assume that the marginal cost of producing the good is zero. In this case, v_1 is measured by the area below the demand function.[56] If the innovation endures forever, its present social value over its entire lifetime is v_1/r, whereby r denotes the discount rate.

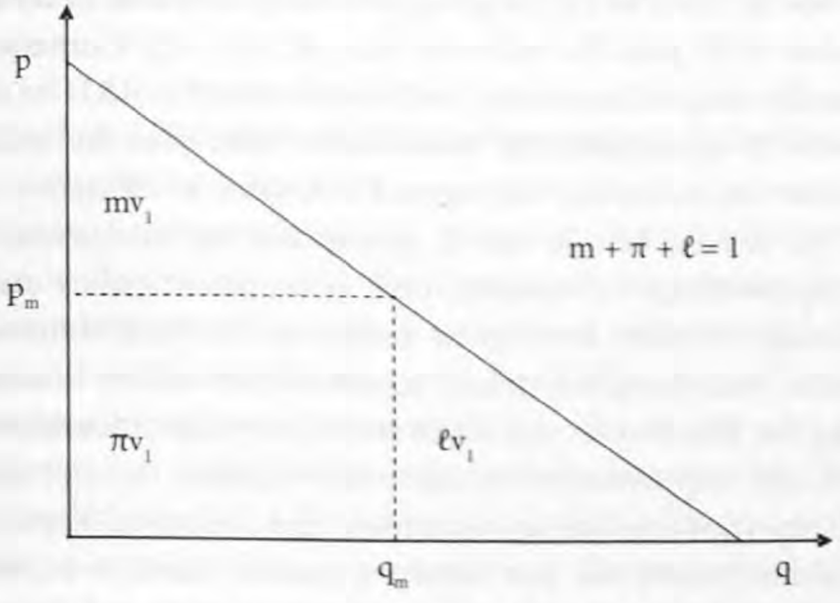


FIGURE 7.1: Per-period Value of Product Innovations

If the innovation is protected by a patent, consumers will only appropriate part of the social value. During the life of the patent, the consumers' share of the innovation's per-period value is only mv_1, with $m < 1$; another part of the social value—πv_1, with $\pi < 1$—is appropriated by the patent owner as per-period profit, which is perceived as the return on the R&D investment. The social cost of

[55] S. Scotchmer (2004), *Innovation and Incentives*, Cambridge: MIT Press.

[56] See for example, ibid., p. 33.

inducing innovation by patent protection is the deadweight loss ℓv_1 per period. Consequently, the present social value of the patented innovation is $v_1/r - v_1\ell T$, whereby T denotes the (discounted) time of patent duration.[57] If the patent holder acts as a perfectly discriminating monopolist he will—during the lifetime of the patent—appropriate the whole per period value of the innovation, so that there is no deadweight loss anymore ($\pi = 1$, $m = \ell = 0$).

Assume furthermore that a secondary innovator, inspired by the original one, invests c_2 in R&D in order to improve the original innovation, and thereby increases the value of the new product to all potential users by v_2 per period and its present social value by v_2/r. We assume, for the sake of simplicity, that the improving innovation emerges with certainty and without any delay. Since the improvement is only possible if the original innovation is made, the (maximal) present social value of the original innovation consists of its stand-alone value (v_1/r) plus the spill-over max $\{0, v_2/r - c_2\}$. Consequently, to induce the original innovator to efficiently invest in R&D he should be allowed to appropriate the stand-alone value plus the spill-over value. However, to induce the second innovator to efficiently invest in R&D he should be allowed to appropriate the total social value of the improvement. Obviously, there is no patent policy available that provides efficient investment incentives for both innovators.[58] An optimal patent regime should minimize the welfare losses from distorting the first and second innovator's investment incentives.[59]

When an improvement satisfies the normal requirements of patentability (novelty, non-obviousness, and industrial application) the secondary innovator can obtain a patent. Here, two cases are possible. If the secondary innovator is entitled to use his patent

[57] That means T is calculated in a way that $v_1\ell T$ denotes the present value of the periodical deadweight losses during the lifetime of the patent. See Scotchmer (2004), p. 59.

[58] With patent protection, and the absence of perfect price discrimination by the patent owner, the maximal present social value is not achievable any more since the deadweight loss has to be taken into account. Thus, the actual present social value of both innovators reduces to $(v_1 + v_2)(1/r) - \ell T)$. The balancing of the innovation incentives between primary and secondary innovator is a function of the design of the patent law which allocates total profit $\pi T(v_1 + v_2)$ to both innovators.

[59] See the elaborated model by Chang (1995) who takes the uncertainty of the first innovator over the value and the cost of the second innovator into account.

without obtaining a license from the primary innovator, that is, if he owns a *competing patent*, he would be enabled to appropriate part of the stand-alone value of the original innovator, thereby undermining his incentive to invest in R&D.[60] If the secondary innovator can use his patent only with the consent of the primary innovator, that is, if he owns a *subservient (dependent) patent*, both parties can block each other and the primary innovator may appropriate part of the value created by the secondary innovation.[61]

Basically, there are two policy levers available to tailor both parties' incentives by patent law: the requirement for patentability of the secondary innovation, that is, for example, the distinction between (non-patentable) 'minor improvements' and (patentable) 'significant or radical improvements',[62] and the breadth of the original patent, that is, the distinction between improvements that infringe or that do not infringe on the original patent. Consequently, we can distinguish four cases:

Case 1: Broad Protection of Original Innovation, Patent for Improving Innovation

In this case, both innovators hold a patent, but the secondary innovator can use his patent only with the consent of the primary innovator. The secondary innovator owns a *subservient (dependent) patent* and is only able to market the improved good if he buys a license from the original innovator.[63] Let us start with the simple case of an *ex post license*, that is, the bargaining about the license takes place after the secondary innovator has invested c_2 in R&D. Since c_2 is sunk, the bargaining surplus is the present value of the periodical profits the secondary innovator earns by selling the improved product: $v_2\pi T$. According to the Nash bargaining solution, each party to the license agreement receives the amount of the threat point, that is, the pay-off without cooperation, plus half of the bargaining surplus. Thus, ex post licensing leads to the following result:

[60] See Ibid. (1995, 44 f.).

[61] Ibid.

[62] M.A. Lemley (1997), 'The Economics of Improvement in Intellectual Property Law', *Texas Law Review*, vol. 75, p. 1007*ff*.

[63] Scotchmer (2004), p. 135.

	Threat point	*Cooperative pay-off*
First innovator	$\pi Tv_1 - c_1$	$\pi Tv_1 + ½ \pi Tv_2 - c_1$
Second innovator	$- c_2$	$1/2 \pi Tv_2 - c_2$

With ex post bargaining, the first innovator is able to hold-up the owner of the subservient patent whose R&D costs are already sunk. This constitutes a problem when $1/2 \pi Tv_2 < c_2 < \pi Tv_2$, because in this case the secondary innovator has no incentive to engage in some socially valuable R&D activities before agreeing upon a license with the primary innovator.

Basically, this problem could be solved by agreeing on an ex ante license, before the secondary innovator has spent c_2. In this case, the bargaining surplus is $\pi Tv_2 - c_2$, and ex ante licensing will be in the mutual interest of both parties, even if $1/2 \pi Tv_2 < c_2 < \pi Tv_2$. Thus, ex ante bargaining on a license leads to the following results:

	Threat point	*Cooperative pay-off*
First innovator	$\pi Tv_1 - c_1$	$\pi Tv_1 + 1/2 \pi Tv_2 - c_1 - 1/2c_2$
Second innovator	0	$1/2 (\pi Tv_2 - c_2)$

Consequently, more socially valuable improvements will be made if we allow for ex ante licensing. However, this solution might not be feasible because of high transaction costs or because this type of cooperation conflicts with antitrust law.[64]

In general, the combination of broad patent protection for original innovators and patents for improvers creates blocking patents and thereby incentives to disclose new knowledge and to allow rivals to build on the current state of knowledge.

Case 2: Narrow Protection of Original Innovation, Patent for Improving Innovations

In this case, the secondary innovator is allowed to sell the goods that embody the secondary innovation without asking the primary innovator for a license ('competing patents'). This enables the secondary innovator to compete with the original innovator, thereby lowering his profits

[64] H.F. Chang (1995), 'Patent Scope, Antitrust Policy, and Cumulative Innovation', *RAND Journal of Economics*, vol. 26, pp. 34–57.

from the original innovation. Instead of earning the monopoly profit πTv_1 during the lifetime of the patent, he will only earn the oligopoly profit $\pi' Tv_1$, with $\pi' < \pi$. Consequently, narrow patents for original innovations discourage investment in R&D by primary innovators and induce them to delay the disclosure of their innovation until they have improved the original innovation themselves.

From this reasoning it follows that there are two classes of primary innovations that should receive broad patent protection in order to efficiently trade-off between primary and secondary innovator's incentives: First of all, those basic innovations which have a high stand-alone value relative to possible improvements (which is intuitively clear). Second, basic innovations with a very little stand-alone value relative to the high value of expected improvements. The reason for a broad protection in the second case is the existence of large spillovers that are created by the basic innovation and that require a strong incentive for the primary innovator to invest in R&D. There is a special need to protect primary innovations with little stand-alone value when there is a large combined welfare effect resulting from both the primary and secondary innovation.[65] If $v_2(1/r - \ell T) > 0$ and $v_1(1/r - \ell T) + v_2(1/r - \ell T) - c_1 - c_2 > 0$, but $\pi' Tv_1 - c_1 < 0$, competing patents will discourage investment in the original innovation, thereby also destroying the options for secondary innovations. Broad patent protection of primary innovators will enable them to appropriate part of the spillover via licensing, encourage investment in R&D, and see society better off.

This result contradicts the 'reverse doctrine of equivalents' which is applied by some US courts and also supported by Merges and Nelson.[66] 'Under that doctrine, if the contribution made by the improvement *greatly* exceeds the contribution made by the original patented invention, the improver is allowed to practice his invention without being deemed an infringer, even though he is making use of the prior invention without a license from the patentee.'[67]

[65] R. Cooter and T. Ulen (2004), *Law and Economics*, 4th edition, Boston: Pearson Education, p. 126.

[66] R.P. Merges and R.R. Nelson (1990), 'On the Complex Economics of Patent Scope', *Columbia Law Review*, vol. 90, pp. 839–916.

[67] W.M. Landes and R.A. Posner (2003), *The Economic Structure of Intellectual Property Law*, Cambridge: Harvard University Press.

If 'minor improvements' are patentable, more general problems may arise. There might be a tendency to patent small improvements too early, thereby depriving potential future innovators of the opportunity to develop a better innovation.[68] If many minor improvements without a remarkable value added are protected by patents, a patent thicket[69] would be created which would require developers to negotiate with many patent holders and discourage companies from developing more valuable improvements. Also, transaction costs of licensing increase and the 'tragedy of the anticommons'[70] may arise.

Now, assume that secondary innovations are not patentable since they do not meet the requirements of patentability. This corresponds to Section 3(d) of the Indian Patent (Amendment) Act 2005 which determines specific restrictions on patenting secondary innovations.

Case 3: Broad Protection of Original Innovation, No Patent for Improving Innovation

This constellation discourages investments in the improvement of original innovations and also the disclosure of improving innovations in due time. The secondary innovator's best strategy is to wait and hide the information, and (if it is profitable) to continue with R&D until a level of improvement is achieved which makes the product patentable. Consequently, technical progress is retarded and investment in R&D for primary innovations with large expected spillovers might be discouraged.

Case 4: Narrow Protection of Original Innovation, No Patent for Improving Innovation

When the improving innovation is not patentable, but does not infringe on the original patent, the secondary innovator faces two options:[71] The first option is to market the good which embodies the

[68] Denicolò (2008), p. 444.

[69] C. Shapiro (2001), 'Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting', in A. Jaffe, J. Lerner and S. Stern (eds), *Innovation Policy and the Economy*, vol. I, Cambridge, pp. 119–50.

[70] M.A. Heller and R.S. Eisenberg (1998), 'Can Patents Deter Innovation? The Anticommons in Biomedical Research', *Science*, vol. 280, pp. 698–701.

[71] Scotchmer (2004, p. 150.

non-patentable improvement, provided it can be protected by trade secrecy or first-mover advantages. In the latter case, some of the new knowledge may be revealed to the public, but the original innovator's profits will be eroded, which discourages his investment in R&D in the first place. The second option is to cache the improvement and wait for further (patentable) improvements (see case 3).

Will there be any change when the original and the incremental innovation are provided by the same party? Since all spillovers are internalized, the problem of coordinating the R&D activities between primary and secondary innovators vanishes. But proponents of a restrictive policy with respect to improvements of patented innovations, such as the creators of Section 3(d) of the Indian Patent Act, argue that another problem will arise: 'evergreening'. Evergreening generally refers to the strategy adopted by patentees who seek to extend their period of patent protection by applying for secondary patents over related or derivative technologies.[72] However, does evergreening really constitute a practical problem? Imagine that a pharmaceutical company invents a new drug which is patented so that the residual lifetime of the patent after approval of the drug by the regulatory authority is, for example, fifteen years. After some time, it develops a minor improvement of this drug which is also patented, even though it does not create any remarkable value added, so that the residual lifetime of the new patent after the expiry of the original patent is ten years. Does the new patent extend the patent protection of the original innovator? Obviously not. After the original patent has expired and after the producers of generics have gained access to the relevant product data disclosed by the original innovator to the regulatory authorities the imitators will be entitled to produce generics of the original drug and sell them at a low price. Since the original drug and the newer ones with minor improvements are close substitutes, the price of both drugs will approach the marginal cost of production. This implies that no extension of the original patent and no increase of the deadweight loss is possible.

[72] R. Chambers (2006), 'Evergreen or Deciduous? Australian Trends in relation to the "Evergreening" of Patents', *Melbourne University Law Review*, vol. 30, pp. 29–61.

Only under very special circumstances will the patenting of minor improvements lead to 'evergreening' of the original patent: (i) If the producer of a generic from the original drug infringes on the claims related to the new (improvement) patent. However, in this case the producers of generics would challenge the new patent (which does not comply with the requirements of novelty and/or non-obviousness). (ii) When the pharmaceutical company is able to convince doctors and health insurance systems that the improved drug is much better than the original even though they are in fact close substitutes. However, from our point of view, there are no convincing arguments that this case will exist. In high income countries, there are elaborated checks and balances to keep health costs down, so that the health insurance systems would refuse to cover high-price-drugs that could easily be replaced by low-price-drugs of the same quality. In low income countries, people simply could not afford to pay the high prices.

Of course, the original innovators are typically also very innovative with respect to delaying the entry of generic drugs.[73] However, this does typically not result from the overprotection of marginal improvements, but to a large extent from infringements of antitrust law.

7.5 IMPACT OF RESTRICTIVE INTERPRETATION OF LAW ON INNOVATION

At first glance, the restrictive interpretation of Section 3(d) of the Indian Patent (Amendment) Act 2005 by the Madras High Court, which identifies 'efficacy' with 'therapeutical efficacy', seems to reflect the old conflict between producers of original drugs, located in developed countries, and producers of generics, located in developing countries. However, this is not necessarily true. As it has become clear from the economic analysis, the narrow interpretation of Section 3(d) severely restricts innovation through incremental steps, and this restriction holds for pharmaceutical companies from developed countries and from India as well.

[73] In July 2009, the European Commission presented the final report on its competition inquiry into the pharmaceutical sector. It found out 'that originator companies use a variety of instruments to extend the commercial life of their products without generic entry for as long as possible.' http://europa.eu/rapid/press ReleasesAction.do?reference=IP/09/1098&format=HTML (last accessed on 17 July 2009).

There is some evidence that India has developed a huge innovative potential in the field of pharmaceutical research and development. In the last years more and more Indian pharmaceutical companies have spent larger fractions of revenues in research and development (R&D). Whereas in 1992–1993 seven major spenders invested only 1.78 per cent of sales in R&D, the investment of twenty-eight major spenders amounted to 8.79 per cent of sales in R&D in 2005–6.[74] Some big pharmaceutical companies like Ranbaxy Laboratories and Dr Reddy Laboratory Ltd. have now started taking R&D seriously (see the corresponding Annual Reports).

Under these conditions, the restrictive interpretation of Section 3(d) also works against the Indian pharmaceutical companies that have the potential to make existing drugs more useful for patients under the specific conditions of countries with tropical climates, high rates of illiteracy, etc.

REFERENCES

Banbury, C.M. and W. Mitchell (1995), 'The Effect of Introducing Important Incremental Innovations on Market Share and Business Survival', *Strategic Management Journal*, vol. 16, pp. 161–82.

Basheer, S. (2008), 'The "Glivec" Patent Saga: A 3-D Perspective on Indian Patent Policy and TRIPS Compliance. Paper presented at the Indo-German Conference on Intellectual Property Law', Freiburg and Munich, 12–15 May 2008, available at http://www.atrip.org/upload/files/essays/Shamnad%20Basheer%20Glivec%20Patent%20Saga.doc (last accessed on 25 May 2009).

—— (2005), 'Limiting the Patentability of Pharmaceutical Inventions and Micro-Organisms: A TRIPS Compatibility Review', Mimeo.

Basheer, S. and T.P. Reddy (2008), 'The "Efficacy" of Indian Patent Law: Ironing out the Creases in Section 3(d)', *Scripted*, 5(2), pp. 232–66, available at http://ssrn.com/abstract=1086254 (last accessed on 25 May 2009).

Benkard, G., (2006), *Patentgesetz*, Munich: S. Beck.

Bently, L. and B. Sherman (2001), *Intellectual Property Law*, New York: Oxford University Press.

Bhaskaran, S. (2006), 'Incremental Innovation and Business Performance: Small and Medium-Size Food Enterprises in a Concentrated Industry Environment', *Journal of Small Business Management*, vol. 44, pp. 64–80.

[74] Chaudhuri (2007), p. 22.

Burdon, M. and K. Sloper (2003), 'The Art of Using Secondary Patents to Improve Protection', *International Journal of Medical Marketing*, 3(3), pp. 226–38.

Cardinal, L.B. (2001), 'Technological Innovation in the Pharmaceutical Industry: The Use of Organizational Control in Managing Research and Development', *Organizational Science*, vol. 12, pp. 19–36.

Chambers, R. (2006), 'Evergreen or Deciduous? Australian Trends in Relation to the "Evergreening" of Patents', *Melbourne University Law Review*, vol. 30, pp. 29–61.

Chang, H.F. (1995), 'Patent Scope, Antitrust Policy, and Cumulative Innovation', *RAND Journal of Economics*, vol. 26, pp. 34–57.

Chaudhuri, S. (2007), 'Is Product Patent Protection Necessary in Developing Countries for Innovation? R&D by Indian Pharmaceutical Companies after TRIPS', Working Paper No. 614/September 2007, Indian Institute of Management, Calcutta.

Cool, K. (1985), 'Strategic Group Formation and Strategic Group Shifts: A Longitudinal Analysis of the U.S. Pharmaceutical Industry', 1963–1982, Unpublished Doctoral Dissertation, Purdue University, West Lafayette, IN.

Cooter, R. and T. Ulen (2004), *Law and Economics*, 4th edition, Boston: Pearson Education.

Denicolò, V. (2008), 'Economic Theories of the Nonobviousness Requirement for Patentability: A Survey', *Lewis & Clark Law Review*, vol. 12, pp. 443–59.

Dewar, R.D. and J.E. Dutton (1986), 'The Adoption of Radical and Incremental Innovations: An Empirical Analysis', *Management Science*, vol. 32, pp. 1422–33.

Ducor, Philippe S. (1998), 'Patenting the Recombinant Products of Biotechnology and Other Molecules', The Hague: Kluwer Law International, p. 30.

Duchesneau, T.D., S. Cohn, and J. Dutton (1979) 'A Study of Innovation in Manufacturing, Determination, Processes and Methodological Issues', Vol-I Social Science Research Institute, University of Maine.

Ettlie, J.E. (1983), 'Organizational Policy and Innovation among Suppliers to the Food Processing Sector', *Academic Management Journal*, vol. 26, pp. 27–44.

Falvey, R.E., F. Martinez, and G.V. Reed (2002), 'Trade and the Globalisation of Patent Rights', GEP Research Paper No 02/21, available at http://ssrn.com/abstract=413263 (last accessed on 10 July 2009).

Green, J.R. and S. Scotchmer (1995), 'On the Division of Profit in Sequential Innovation', *RAND Journal of Economics*, vol. 26, pp. 20–33.

Grubb, Ph. W. (2004), *Patents for Chemicals, Pharmaceuticals and Biotechnology: Fundamentals of Global Law, Practice and Strategy*, 4th edn, Oxford: Oxford University Press.

Heller, M.A. and R.S. Eisenberg (1998), 'Can Patents Deter Innovation? The Anticommons in Biomedical Research', *Science*, vol. 280, pp. 698–701.

Hoekmann, B. and M. Kostecki (2001), *The WTO and Beyond*, 2nd edn, Oxford: Oxford University Press, p. 274.

Jacob, R. (1996), *Novelty of Use Claims*, 27 IIC, pp. 170–81.

Kamperman Sanders, A. (2007), 'Intellectual Property Law and Policy and Economic Development with Special Reference to China', in T. Eger, M. Faure, and Z. Naigen (eds), *Economic Analysis of Law in China*, Cheltenham, pp. 239–71.

Landes, W.M. and R.A. Posner (2003), *The Economic Structure of Intellectual Property Law*, Cambridge: Harvard University Press.

C. Langinier, G. Moschini (2002), *The Economics of Patents: An Overview*. Iowa State University, Center for Agricultural and Rural Development, Working Paper 02-WP 293.

Lemley, M.A. (1997), 'The Economics of Improvement in Intellectual Property Law', *Texas Law Review*, 75, pp. 989–1084.

Lévêque, D. and Y. Ménière (2004), *The Economics of Patents and Copyright*, Paris, available under http://www.cerna.ensmp.fr/Documents/FL-YM-eBookIP.pdf (last accessed on 25 May 2009).

Levy, R.A. (1990), 'Pharmaceutical Research: Therapeutic and Economic Value of Incremental Innovations', Unpublished Dissertation, National Pharmaceutical Council, Reston, VA.

Mansfield, E., (1986), 'Patents and Innovation: An Empirical Study', *Management Sciene*, 32(2), pp. 173–81.

Merges, R.P. and J.F. Duffy (2007), *Patent Law and Policy*, 4th edn, Newark, NJ: LexisNexis.

Merges, R.P. and R.R. Nelson (1990), 'On the Complex Economics of Patent Scope', *Columbia Law Review*, vol. 90, pp. 839–916.

Min, S., M.U. Kalwani, and W.T. Robinson (2006), 'Market Pioneer and Early Follower Survival Risks: A Contingency Analysis of Really New Versus Incrementally New Product-Markets', *Journal of Marketing*, vol. 70, pp. 15–33.

Schulte, R. (2001), *Patentgesetz*, 6th edn, Cologne: Heymanns.

Scotchmer, S. (2004), *Innovation and Incentives*, Cambridge: MIT Press.

Shapiro, C. (2001), 'Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting', in A. Jaffe, J. Lerner, and S. Stern (eds), *Innovation Policy and the Economy*, vol. I, Cambridge, pp. 119–50.

Shavell, S. (2004), *Foundations of Economic Analysis of Law*, Cambridge: Harvard University Press.

Thomas, J.J. (2006), 'India and China in the Knowledge Economy: Rivals or Allies? Case Studies of Pharmaceuticals and Biotechnology', Paper presented at the Conference on China, India, and the International Economic Order, Singapore, 23–4 June.

Thomas, J.R. (2005), *Pharmaceutical Patent Law*, Washington DC: BWA Books.

Trask, A.V. (2008), 'Obvious To Try: A Proper Patentability Standard in the Pharmaceutical Arts?' *Fordham L. Rev.*, vol. 76, p. 2625.

Van den Bossche, P. (2007), *The Law and Policy of the World Trade Organisation*, Cambridge: Cambridge University Press.

Watal, J. (2008), 'Patents', in K. Basu (ed.), *The Oxford Companion to Economics in India*, rev. edn, Delhi: Oxford University Press, pp. 400–3.

Zaltmang Gerald, Robert Duncan, and Jonny Holbek (1973), *Innovations and Organizations*, New York: John Wiley & Sons.

IV

CORRUPTION

Shapiro, C. (2001), 'Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting', in A. Jaffe, J. Lerner, and S. Stern (eds), *Innovation Policy and the Economy*, vol. 1, Cambridge, pp. 119–50.

Shavell, S. (2004), *Foundations of Economic Analysis of Law*, Cambridge: Harvard University Press.

[illegible], C. (2006), 'India and China in the Knowledge Economy: the Case of Albert Case Studies of Pharmaceuticals and Biotechnology', Paper presented at the Conference on China, India, and the International Economic Order, Singapore, 23–4 June.

Thomas, J.R. (2005), *Pharmaceutical Patent Law*, Washington DC: BNA Books.

[illegible], A.V. (2008), 'Obvious To Try: A Proper Patentability Standard in the Pharmaceutical Arts?' *Fordham L. Rev.*, vol. 76, p. 2625.

Van den Bossche, P. (2005), *The Law and Policy of the World Trade Organization*, Cambridge: Cambridge University Press.

Watal, J. (2010), 'Patents', in K. Basu (ed.), *Oxford Companion to Economics in India*, new edn, New Delhi: Oxford University Press, pp. 400–3.

Zaltman, Gerald, Robert Duncan, and Jonny Holbek (1973), *Innovations and Organizations*, New York: John Wiley & Sons.

8

ENFORCEMENT AND CORRUPTION

Some Implications of 'Soft' Information

Ajit Mishra*

8.1 SOFT INFORMATION AND CORRUPTION

Corruption is widely perceived to be detrimental to economic development. The recent cross-country empirical studies[1] have shown a predominantly negative link between perceived levels of corruption and various macro-economic variables like growth, investment, and inequality. Though the moderate-to-successful growth experience of many countries with high levels of corruption along with the difficulties in establishing causal relationships have raised quite a few questions about these studies, the link between corruption and low levels of development can hardly be denied. If anything, the link operates both ways—poorer countries also happen to be (or are perceived to be) more corrupt. In this paper, we take a more micro-perspective and look at the implications of certain factors like the nature of the information and the judicial access to corruption in the context of developing countries. In keeping with the theme of this workshop, we focus on corruption in the specific context of regulatory and legal enforcement. Enforcement

* I would like to thank the participants at the Asia-Link Senior Scholar Workshop on 'Law and Economic Development,' Hyderabad, February 2008. I have also benefited from comments by a referee.

[1] Basically these look at correlations (with implications for causation) between economic performance indicators and some measure of country-level corruption across a set of countries. For example, looking at growth performances of different countries, we could ask whether a country is likely to grow slowly because its corruption level is high. See Mishra (2009) for a survey.

problems have been studied by many scholars in the extensive and rich law and economics literature. Our objective here is to examine how the 'soft' nature of information and the relatively costly nature of judicial access affect corruption and enforcement policies.

Information can be viewed as 'soft' if it cannot be verified easily (without incurring significant costs) and can consequently be manipulated easily. In a developing country, for various reasons, information tends to be soft in many settings. For example, in a developed country one is likely to get a speeding ticket only when there is some photographic evidence (though it could be misleading!) while in many developing countries the speeding ticket could simply be based on somebody's observation. Similarly, the media and various investigating agencies—typical sources of information—tend to lack independence and credibility in many of the developing countries. Coupled with this, verification costs are also high. In particular, access to courts and similar bodies is costly, especially when procedural delays and hurdles are taken into consideration. These two factors are hardly mentioned in many discourses on corruption; it is our objective to draw attention to these and look at some of the implications.

Corruption is commonly defined as the misuse of public office or authority for private gains.[2] The three terms 'misuse', 'public' and 'private gain' are subject to various interpretations and lead to alternative formulations. The multi-dimensional and context-specific nature of corruption makes it hard to define in an inclusive manner. It is no surprise that researchers have analysed corruption in different manner. We choose to focus on the problem of enforcement, and discuss corruption in this particular context. Corruption dilutes and distorts enforcement policy. Given the context, our modelling of corruption relies on previous research in the law and economics tradition.[3] Specifically, we use the agency framework to address issues related to corruption.

[2] Most papers in economics tend to use this definition. It is also used by multilateral organizations such as UNDP (2004), *Anti-corruption. Practice note*. New York: UNDP. World Bank (2007), *Strengthening World Bank Group Engagement on Governance and Anti-Corruption*, Washington DC: World Bank.

[3] See, for example, 'The Public Enforcement of Law', Polinsky and Shavell (2000).

In an agency setting, a law or regulatory standard or policy (for a set of clients) is enforced by the authority through a set of intermediate agents who are in charge of collecting and reporting relevant information (that is, detection of non-compliance). Corruption arises because of the ability and scope of these agents to manipulate such information. The assumption of soft information essentially increases the power of the agents in manipulating information. We explore two key implications of this feature.

In the first instance we consider an example where information acquisition is costly but the agent could effectively cook up information. This leads to avoidable compliance costs for the clients. If judicial redress is costly or unavailable, such compliance costs are likely to be higher. We show that even when there is an incentive system in place to discourage such behaviour, society could get stuck in a low-compliance trap. Secondly, soft information makes it harder to implement high-powered incentive schemes in the context of enforcement. We consider an example of marginal deterrence where soft information and costly judicial redress makes under or over enforcement more likely. Under enforcement refers to the case where some less harmful activities are legalized. Over enforcement implies that some clients are prevented from choosing activities even when their private benefits exceed social costs.[4]

In Section 8.2 we present the basic agency framework. Section 8.3 looks at the various anti-corruption strategies proposed within this framework. Apart from providing a brief review of the common strategies, we wish to focus on their informational requirements and to draw attention to the important role of information in fighting corruption. Section 8.4 analyses the two main implications of soft information that we have highlighted earlier.

8.2 CORRUPTION IN AGENCIES

Analysing corruption in an agency framework is useful because it enables us to identify the various possible sources of corruption. An

[4] The issue of over enforcement has received some attention by legal scholars. See R.A. Bierschbach and A. Stein (2005), 'Overenforcement', *The Georgetown Law Journal*, vol. 93, pp. 1743–81, and T.F. Cotter (2006), 'Fair Use and Copyright Overenforcement', *Legal Studies Research Paper* Series 6-69, University of Minnesota Law School.

agency relationship arises when two individuals enter a non-market (contractual) relationship and where one individual (commonly termed as the principal) relies on another individual (commonly referred to as the agent) to carry out certain actions on his behalf. The actions of the agent affect the principal's payoff in a significant way. Problems arise in such relationships when (i) the principal and the agent have dissimilar objectives and (ii) the principal is unable to enter into a comprehensive enforceable contract with the agent regarding the desired course of action.

Corruption involves more than one overlapping relationship. In the government sphere, the first is the one between the government and its various clients. For example, relationships between government and tax payers, regulator and firms, police authority and potential criminals, government providers of services and potential recipients can all be viewed as principal-client relationships. The identities of the principals and the clients as well as the nature of their relationships vary depending on the specific context. Thus, potential criminals who hope to avoid detection when they break the law are very different from those who seek to obtain government benefits. The scope for corruption arises when principals employ agents to govern their relationships with the various clients. When agents deviate from their contractual (explicit or implicit) obligations to their principals and act in a way that serves their own interests, it will be considered a corrupt act. In most cases such acts are likely to undermine the interest of the principal in the specific principal–client relationship. Clearly, it covers various forms of corruption.

To fix ideas concerning the underlying agency relation, consider the example of a regulatory authority trying to enforce pollution standards by using pollution inspectors who are in charge of inspecting the polluting firms. The early work by Susan Rose-Ackerman (1978) focuses mainly on the agency relationship between the authority and the inspector. Corruption arises when 'some third person, who can benefit by the agent's actions, seeks to influence the agent's decision by offering him a monetary payment which is not passed on to the principal'.[5] In our example, a polluting firm is the third person. Robert

[5] Susan Rose-Ackerman (1978), *Corruption: A Study in Political Economy*, New York: Academic Press.

Klitgaard[6] uses a similar framework where the principal is the regulator, the agent the inspector, and the client is the firm. In economics, Jean Tirole's (1986) early work on collusion in organizations using the principal–supervisor–agent relationship has led to a substantial literature on collusion, some of which dealt with corruption.

We can classify relationships according to the extent of power and responsibilities enjoyed by the agent. Informational asymmetry and contractual incompleteness are the two main sources of such power. First, the agent may have a purely information-gathering role. In such a case, power will come from the ability to manipulate such information. Most enforcement (regulatory enforcement, tax enforcement, policing) problems would fall into this category.[7] As stated earlier, our paper primarily analyses this kind of corruption. Second, the principal might set some broad objectives for the agent who would have the power to choose the exact incentive mechanism for the clients. High-level bureaucratic corruption in policy design and delivery of public services would resemble this scenario. The bureaucrat may have a rough target (so many poor people to receive the public services, so much of revenue to be generated) but otherwise has full control. Third, the principal might simply transfer all the power to the agent. Issuing of licenses, permits, etc., where guidelines are virtually non-existent or non-operational resemble this monopoly case.[8]

8.3 ANTI-CORRUPTION STRATEGIES

We shall use a simple model of enforcement within the principal-agent–client framework to study the various components of anti-corruption strategies. As mentioned earlier, we have a group of clients

[6] R. Klitgaard (1988), *Controlling Corruption*, Berkeley and Los Angeles: University of California Press.

[7] J. Tirole (1986), 'Hierarchies and Bureaucracies: On the role of Collusion in Organizations', *Journal of Law, Economics and Organizations*, vol. 2, pp. 181–214, F. Kofman and J. Lawarree (1993), 'Collusion in Hierarchical Agency', *Econometrica*, vol. 61, pp. 629–56. D. Mookherjee and I. P'ng (1995), 'Corruptible law enforces: how should they be compensated?' *Economic Journal*, vol. 105, pp. 145–59, Polinsky and Shavell (2000) are some of the models of corruption in this context.

[8] There is a large literature looking at the inefficiencies resulting from corruption by the monopoly provider (Shliefer and Vishny).

(individuals/firms) who are considering undertaking an illegal activity. These clients are monitored by agents who are supposed to report the activity to the principal. This basic hierarchical relation is shown below. The solid arrows (facing down) refer to the monitoring relation, while the dotted arrow denotes reporting. Since the agents are corruptible, they can always take a bribe and manipulate the report. Hence the principal faces the double incentive problem—of designing incentive mechanisms for both the clients and the monitoring agents.



FIGURE 8.1: Hierarchical Relation among Anti-Corruption Officers

Source: Author.

In keeping with our pollution control scenario mentioned earlier, the group of clients are the firms who are required to adhere to certain pollution norms and the agent is pollution inspector. A typical firm's gain from not complying with the standard is g, and the penalty after being reported by the inspector is f. The principal has to ensure that the inspectors report truthfully.

Corruption is unique because the mechanism designed to fight corruption can itself be compromised. To the extent that corruption is a manifestation of rational economic calculation by public officials with power and discretion, it can be addressed by suitable incentive schemes. These schemes revolve around a basic carrot and stick policy of rewarding the honest and punishing the corrupt. However, these schemes require both successful detection of corruption and proper implementation of incentive schemes. If officials in charge of detection and implementation are also corrupt, then we are forced back to the original problem. Hence fighting corruption requires a different focus

compared to numerous other enforcement problems (that is, crimes, drug trafficking, and money laundering).

8.3.1 Higher Level Monitoring and External Audit

To monitor the key source of information and to prevent mis-reporting by the agent, one can add layers of higher level monitoring in the basic hierarchy described earlier. The agents are monitored by another set of agents and so on. Such hierarchical monitoring and supervision helps but it is costly and the fact that corruption/collusion could prevail at the higher levels makes corruption deterrence more problematic.[9]

In some cases, the delegation and use of external bodies have proved helpful. For example, the practice of pre-shipment inspection schemes (PSI) followed by many countries to combat corruption in the collection of customs and import duties can be viewed as successful in most cases. The PSI scheme was introduced by Phillipines in 1987 where shipments were to be inspected by private inspectors before their departure from the exporting countries. Since the PSI scheme generated an independent source of information about import duty, it certainly made evasion and corruption more difficult. Yang[10] analyses the experience of these countries and concludes that import revenues went up by 8–10 per cent because of these PSI schemes.

Irrespective of whether we use internal or external monitoring, it is clear that the success of the anti-corruption policy crucially depends on the nature of information gathered and reported by the anti-corruption officials.

8.3.2 Rewards and Incentives

Instead of using higher level monitoring to deter corruption by the agent, one can create positive incentives for the agent to be honest. For example, a number of countries have tried various incentive reforms in revenue collection agencies to fight corruption by tax collection

[9] A. Mishra (2006), 'Corruption, Hierarchies and Bureaucratic Structures,' in S. Rose-Ackerman (eds), *International Handbook on the Economics of Corruption*. Chetenham, UK: Edward Elgar contains a discussion of these hierarchical designs and their implications for corruption control.

[10] D. Yang (2008), 'Integrity for Hire: An Analysis of a Widespread Customs Reform', *Journal of Law and Economics*, forthcoming.

agents. This involves a combination of rewards, bonuses, and efficiency wages. The evidence for such incentives is quite mixed; though there was some success in the initial years following the reforms, they have not worked in eliminating corruption. This is partly due to the difficulties in implementing these incentive schemes.[11] First, in many cases enforcement effort is like a joint output. Second, agents could manipulate the reporting of the outcome to claim bogus rewards. Again, the role of information is quite critical here.

8.3.3 Wage Policy and Efficiency Wage

A third way is to combine elements of wage policies and monitoring. Low wages of bureaucrats or civil servants have generally been viewed as a major inducement for corrupt behaviour. There are at least two arguments as to why low wages can push public officials towards corruption. First, when lowly paid officials deal with relatively rich clients, the income inequality and wide difference in living standards could be an inducement for the official to use his/her power in order to climb up the social and income ladder. Second, high wages can act as efficiency wage and this helps deter corrupt behaviour. Since bribe-taking officials run the risk of losing their jobs if caught, high wages effectively work as fines for bribe taking.[12]

The empirical evidence on this issue is quite mixed.[13] But it is clear

[11] See S.K. Chand and K. Moene (1999), 'Controlling Fiscal Corruption', *World Development*, 27(7), pp. 1129–40. O.H. Fjeldstad (2003), 'Fighting Fiscal Corruption: Lessons from the Tanzanian Revenue Authority', *Public Administration and Development*, 23(2), pp. 165–750. R. Taliercio, Jr. (2004), 'Administrative Reforms as Credible Commitment: The Impact of Autonomy on Revenue Authority Performance in Latin America', *World Development*, 32 (2), pp. 213–32. Dilip Mookherjee (1997), 'Incentive Reforms in Developing Country Bureaucracies: Lessons from Tax Administration', *Annual Bank Conference on Development Economics*, World Bank, DC, contains an excellent discussion of the problems associated with implementing such schemes.

[12] The efficiency arguments go back to Becker and G. Stigler Becker and G. Stigler (1974), 'Law Enforcement, Malfeasance and Compensation of Enforcers', *Journal of Legal Studies*, vol. 3, pp. 1–18.

[13] C. Van Rijckeghem and B. Weder (2001), 'Bureaucratic Corruption and the Rate of Temptation: Do Wages in Civil Services Affect Corruption, and by How Much?' *Journal of Development Economics*, 65(2), pp. 307–31, find a negative relation between public sector wages and the level of corruption. D. Triesman, (2000), 'The Causes of Corruption: A Cross National Study', *Journal of Public*

that efficiency wage relies heavily on the ability to monitor. If the probability of detection (of bribery) is small, then high wages will not work as efficiency wages and officials will continue to be corrupt. In fact high wages would possibly increase their bargaining power and raise the level of bribe payments.[14] This could possibly account for the mixed results on the link between wages and corruption. The study by Di Tella and Schargrodsky[15] highlights the role of monitoring in this context. They show that the success of efficiency wage type arguments depend on the auditing intensity.

8.3.4 Monitoring: Top–Down or Bottom–Up?

The hierarchical model studied earlier can be viewed as an example of top-down monitoring. We can have other types of monitoring too. Recently, a number of scholars have emphasized the role of participatory monitoring. In the latter case, a subset of the clients will have to undertake some monitoring. Unfortunately, there is no general result regarding the effectiveness of either method. In the context of corruption, as argued at the beginning, top-down monitoring may not work because the monitors could be corruptible. Similarly, grass-roots monitoring may not work because of free-rider problem and the possibility of local capture.

Reinikka and Svensson[16] discuss how participatory monitoring proved quite effective in the Uganda school grant scheme.[17] In some cases, activists have been involved in the participatory monitoring process as well. There are many instances of such campaigns in India,

---

Economics, vol. 76, pp. 399–457. J.E. Rauch and P.B. Evans (2000), 'Bureaucratic Structures and Bureaucratic Performances in LDCs', *Journal of Public Economics*, vol. 75, pp. 49–71, do not find strong support for this hypothesis.

[14] See Mookherjee and P'ng (1995).

[15] R. Di Tella and E. Schargrodsky (2003), 'The Role of Wages and Auditing During a Crackdown on Corruption in the City of Buenos Aires', *Journal of Law and Economics*, 46(1), pp. 269–92.

[16] R. Reinikka and J. Svensson (2005), 'Fighting Corruption to Improve Schooling: Evidence from a Newspaper Campaign in Uganda', *The Journal of European Economic Association*, 2(2–3), pp. 259–67.

[17] On the other hand Olken's [Benjamin Olken (2007), 'Monitoring Corruption: Evidence from a Field Experiment in Indonesia', *Journal of Political Economy*, 115(2), pp. 200–49] field experiment in Indonesia (KDP scheme by Indonesian Government) shows that top-down monitoring is more effective.

though these have not been systematically documented. Seghal (2004) gives an account of one such instance. In India, the poorest of the poor category receive entitlement to five kg of wheat at Rs 2 and ten kg of rice at Rs 3 per kg. The open market prices are much higher (Rs 13 and Rs 15). In many instances, the ration shop would record (falsely) a sale of the entitled amount with a forged signature (thumb impression in most cases). Vast quantities of rice and wheat would then find their way to the open market where the shop owner would make huge profits. A poor woman, Triveni Devi, aided by the NGO Parivaratan exposed such a scam in Delhi (the capital city).[18] The estimated leakage from the Public Distribution System (PDS) was worth Rs 300 crores in the city alone.

8.3.5 Empowerment and Access to Information

The above example suggests that empowerment and access to information are key factors in promoting participatory monitoring. Limited evidence[19] suggests that, at least in the case of public service delivery, information campaigns amongst the end users can lead to better monitoring of local officials and reduction in the leakage of funds. Following the findings of a Public Expenditure Tracking Survey (in year 1996) to estimate the leakage of funds in the school capitation grant scheme in Uganda, the government introduced an information campaign and citizen empowerment programme by publishing data on monthly transfers of capitation grants to schools in national and local newspapers. A second survey was carried out in 2002 to assess the impact of this campaign. It turns out that while only 24 per cent of the total yearly grants reached schools on an average in 1995, the average school received more than 80 per cent of its entitlement in 2001. Even after controlling for various other factors, this is a substantial improvement. Further, this has led to a substantial improvement in school enrolment and pupil performances.

In the Indian context, the Right to Information Act has had a similar impact in exposing (deterring) corruption in some places.

[18] R. Sehgal (2004), 'Triveni Devi and 109 other women vanquish corrupt ration shop dealers', *Infochange*, June 2004, http://www.infochangeindia.org/features181.jsp (last accessed on 5 July 2009).

[19] Reinnika and Svensson (2005).

For many years, various anti-poverty programmes (Food for Work, Employment Guarantee Scheme) have proved ineffective because of corruption. The experience has been so bad that when the recently enacted National Rural Employment Guarantee Act was being debated in the public and policy circles, one commentator went on to describe it as a corruption guarantee scheme. In many of these employment guarantee schemes, public funds get siphoned off by fudging the muster rolls. Corrupt officials enter fictitious names in the muster rolls and appropriate the wages. However, as Dreze[20] points out, this practice has been significantly curtailed in areas where active use has been made of the Right to Information Act. It is heartening to note that social audits conducted in various places in the state of Rajasthan find that muster rolls have been accessible and have remained clean.

Additionally, there is no denying that the media has a big role to play in this regard. Brunetti and Weder[21] show that in a cross-section of countries, freedom of press is linked to the level of perceived corruption. But there is always the possibility that press itself is vulnerable to manipulation. In this context the role of other non-print media (e-media, radio, etc. where information flow is faster and it is less manipulable needs to be explored). On the other hand, since information is likely to be 'softer', its strength can be its weakness too.

8.3.6 Informational Requirements

Models generating many of the anti-corruption strategies tend to assume that information is 'hard', that is, invulnerable to manipulation. For example, wage reforms or incentive payments are difficult to implement (or even counter productive) if we do not have access to hard information. The same is true for various monitoring schemes. Theoretically, it is not difficult to design mechanisms that successfully curb corruption. But the experience of anti-corruption strategies tells a different story. Our analysis in the next section suggests that we cannot ignore the informational requirements, and

[20] J. Dreze (2008), 'India's Employment Guarantee Act: Reclaiming Policy Space', mimeo.

[21] A. Brunetti and B. Weder (2003), 'A Free Press is Bad News for Corruption', *Journal of Public Economics*, 87(7), pp. 1801–24.

in particular the soft nature of information, while designing anti-corruption strategies.

8.4 IMPLICATIONS OF 'SOFT INFORMATION'

8.4.1 Low-compliance Trap[22]

Consider a group of firms that are expected to adhere to a certain pollution norm. A firm can choose whether to comply with the prescribed standard or not. The gain/benefit from non-compliance g differs across firms and for simplicity we assume that there are only four types of firms; $g \in \{5, 10, 15, 25\}$. Let us assume that each type is equally likely to flout these norms. The gain g can be interpreted as abatement cost savings to the firm, hence different abatement technologies would account for the difference in gains. Firms compare the expected cost of non-compliance with this gain while making their decisions.

We shall assume that there is a regulatory body in place to enforce the pollution standard (assumed to be zero for convenience). Each firm is going to be inspected with probability μ (assumed to be ½) by a corruptible inspector. The inspector can report the firm to the court or the higher authority as a polluter. We assume the court's technology to be perfect so that the firm's pollution level is revealed in the court with certainty. A firm found guilty by the court is charged a fine $f = 50$ and for limited liability reasons, this is the maximum penalty that can be imposed. We assume that the inspector has a high reward/commission rate and gets 48 as reward if the reported firm is found to have polluted. However, in the absence of any mechanism to discover bribery, this does not guarantee honest reporting. But this means that the minimum bribe that a polluting firm can offer is 48. This in itself will act as deterrence since in most situations the bribe payments by the firm will be an increasing function of both the fine f and reward r. Let $x(f, r)$ be the bribe amount that the polluting firm (upon discovery) has to pay the inspector to escape reporting. If bribery is not feasible, a firm with gain g will engage in pollution if and only if $g > \mu f$. On the other hand, if bribery is feasible, a firm with gain g will pollute if and only if $g > \mu x(f, r)$. This resembles the standard incentive mechanism

[22] This example is based on a similar exercise in Mishra (2006).

that would be suggested to ensure honest reporting by the inspector and compliance by the firm.

We shall be looking at a case where information is soft. In the present case, hard information would imply that the inspector after detecting evidence of pollution can suppress it and report lack of sufficient evidence. But the inspector cannot distort it and report a polluting firm as non-polluting (or vice-versa). The nature of soft information implies that the inspector can report a non-polluting firm as well. In that case, the firm is eventually acquitted by the court, but it incurs a cost $L = 30$. This means that there is an effective appeals process, but it is costly. To discourage the inspector from such misreporting we suppose that the penalty for over-reporting is $d = 20$.

Inspection in itself does not reveal anything, it all depends on whether the inspector has put in the necessary effort or not ($e \in \{0,1\}$). Effort is costly, whenever $e = 1$, it costs the inspector an amount $E = 9$. Effort can be given a very broad interpretation to include investments by the inspector in learning, human capital development, or skill acquisition. It is a measure of the competence of the inspector. Competence may have several connotations, but in our model context it implies that a competent inspector is able to differentiate the polluting firms from the non-polluting firms upon inspection. Such inspectors (who have put in the required effort) will be called *informed*.[23]

We can summarise the situation as a two-stage process. In the first stage, each firm chooses whether to pollute or not. The inspector chooses whether to incur the cost e and be informed. The second stage is a simple bribe game. If the firm and the inspector agree on a bribe (including a bribe of zero), the firm is not reported. Otherwise, the firm is reported. The firms and the inspector work out the expected payoffs in this stage while making decisions in the first stage.

The firm's gain g is not observable to anyone other than the firm, though the distribution is known to all. Neither the inspector nor the other firms can observe a particular firm's choice of pollution level. As discussed earlier, a firm will choose to pollute if and only if its private gain g exceeds some threshold level g^*. This threshold level

[23] We assume that effort e is non-contractible.

will depend on the inspector's choice of effort and the nature of the bribe game.[24]

Now we can have a situation where the inspector chooses $e = 0$ and reports the firm with probability one unless the firm offers a bribe of 48 or more. This means the polluting firm can get away by paying a bribe of 48; hence their expected cost from non-compliance is 24. The non-polluting firms also have an expected cost of 15. All firms with gains exceeding 9 will choose to pollute. Hence 75 per cent of all firms will pollute. Given that only a quarter of the firms choose to comply, the inspector's expected cost from reporting is 20/4 which is less than the cost of being informed (9). This is an equilibrium outcome. When the inspector is expected to be uninformed, more firms would choose to be on the wrong side of the law than otherwise. This, in turn, can make the inspector's decision to choose $e = 0$ optimal. So we might see a greater degree of violations of law with uninformed inspector.

Similarly, when firms expect the inspector to be informed, fewer firms would choose to pollute and the number of non-polluting firms would be likely to be high. This, in turn, makes the inspector's decision to choose $e = 1$ an optimal one. Suppose $e = 1$, then the inspector will never report a non-polluting firm (because of the penalty for over reporting $d > 0$). The expected cost from polluting is 24 but the expected cost from non-polluting is 0. Hence unlike in the previous case, only firms with gain $b > 24$ will choose to pollute. This means that only 25 per cent of all firms choose to pollute. Given that there are many non-polluting firms, the inspector will prefer to be informed. The expected cost of reporting a firm without being informed is 15, which is higher than the cost of being informed.

Let us consider the case of hard information. Now the inspector cannot report anyone without hard evidence. This would require choosing $e = 1$ or being informed. The inspector's choice of $e = 0$ would fetch zero payoff. On the other hand a choice of $e = 1$ would fetch a net expected payoff of 3. Hence the only equilibrium outcome is the high compliance outcome.

[24] Depending on the choice of the inspector, the firms might be facing an informed or uniformed inspector. The non-polluting firms would prefer an informed inspector to an uninformed one. The exact opposite preference is expected for the polluting firms.

Similarly a lower value of L means that judicial remedies are not very costly and the incentive to remain compliant would be high. It can be checked that if $L \leq 26$, the inspector will always choose $e = 1$. When $e = 0$, any firm with a gain exceeding 11 will choose to pollute (it can be checked that the inspector still follows the strategy of reporting a firm unless offered a bribe of 48). That means half the firms would be polluting. But this would make the expected cost of reporting a firm without being informed high compared to the cost of being informed. Hence the low compliance outcome is not an equilibrium outcome. The previous discussion can be summarized in the following proposition.

Proposition: When information is 'soft' and the cost of appeals or verification is high, an economy is more likely to get stuck in the inefficient low-compliance equilibrium.

It can be argued that in the medium term judicial costs (L) cannot be affected, whereas we could stipulate higher fines (f, d). At first glance, it is possible to do so but it may not eliminate the low compliance equilibrium. Recall that in our example, f is assumed to be set at the maximum. It is not possible to raise fines further because of limited liability reasons. Moreover, maximal fines are not necessarily desirable if detection is imperfect and fines could be mistakenly imposed. Similar reasoning applies to the penalty for over reporting. It is possible to come up with examples where a suitable choice of fines could eliminate the low compliance. But our objective is to highlight the fact that we may need to go beyond fines and rewards, and look at the deeper factors like the nature of information and judicial access.

8.4.2 Marginal Deterrence

We consider an extended version of the previous model of enforcement.[25] There are three types of firms, and each firm chooses a pollution level

[25] This is a slightly modified and discrete version of Mookherjee and Png (D. Mookherjee and I. P'ng (1994), 'Marginal Deterrence in Enforcement of Law', *Journal of Political Economy*, vol. 102, pp. 1039–66). They have studied marginal deterrence in absence of corruption. See S. Shavell (1992), 'A Note on Marginal Deterrence', *International Review of Law and Economics*, vol. 12, pp. 345–55 also. The material in this section is based on Mishra (A. Mishra (2007), 'A Model of Enforcement with Collusion and Extortion', mimeo, University of Bath, UK).

which is either zero, medium, or high. A higher pollution level leads to a higher social cost. A firm's private benefit depends on the level of pollution and its own type (for the same level of pollution, one type has higher private benefit). Since the regulator is not privy to information about the firm's type the monitoring intensity cannot be dependent on it. Hence the regulator cannot condition monitoring intensity on the firm's type. Thus, all firms are subject to same rate of monitoring but, the sanctions for different levels of pollution are different. This is a classic case where the enforcement policy could be designed to achieve marginal deterrence. Marginal deterrence refers to a situation where a more harmful activity (high pollution) is deterred because its sanction exceeds that of a less harmful activity.

Consider the benchmark case where the cost of enforcement is zero. Social welfare is the sum of private benefits minus social costs and enforcement costs. Suppose the benefits and costs are such that social welfare is maximized when a type 1 firm chooses no pollution, a type 2 firm chooses a medium level of pollution, and type 3 chooses the highest level of pollution. This is possible, for example, when the type 3 firms' benefit from polluting far outweighs the costs associated with pollution. Such an action profile can be enforced if the monitoring costs are not too high and fines are not constrained to be too small.

But suppose the auditing is done by corruptible inspectors. We assume away the information acquisition problem of the previous case, so the inspector does not have to incur any cost to know the pollution level of the firm being audited. Based on the report of the inspector, the firm will be subject to a fine corresponding to the reported level of pollution.

The regulator may not always implement the optimal benchmark schedule of actions. In particular we shall consider two kinds of cases. The regulator might decide on under-enforcement, where a firm might be allowed to choose an action even if the associated harm exceeds the private benefit. This is like an effective legalization of certain harmful activity. In the second case we might have a policy of over-enforcement where a firm's benefit from an action exceeds the social cost associated with it, and yet the firm is prevented from choosing the particular action. Obviously, complete deterrence is also an example of over-enforcement in this specific example.

Suppose the inspector is corruptible and can manipulate information about the pollution level. As in our previous discussions, hard information would imply that the inspector can suppress information about the pollution level and claim to have found no conclusive evidence (null report) about the level. In our example, this implies that the firm gets away with the smallest possible fine whenever a null report is submitted, irrespective of the actual level of pollution. This generates scope for collusion between the firm and the inspector, the stakes being higher for higher levels of pollution.

Here we have a case of possible collusion between the inspector and the firm. It is clear that enforcement is severely diluted in this case. But this is relatively easy to address. If transfers do not enter our social welfare function, then collusion can be prevented without adding to the enforcement cost by announcing rewards for reports of pollution levels. It is clear that collusion alone does not affect the optimal enforcement levels.

But the result changes when we allow for information to be soft. As discussed earlier, when information is soft the inspector can over-report or under-report. The inspector can report even when he does not observe the true level of action. The analysis of under-reporting is similar to the previous analysis of collusion. However, the possibility of over-reporting affects firms' choice of action in a significant way. To avoid such misreporting we can have a system of appeals, whereby the firm could appeal following the report by the inspector. This appeal process is costly and we assume that the firm's action is revealed with certainty. Like the previous example, if there is a positive cost for the inspector following a successful appeal by the firm, the inspector will not over-report if the firm is going to appeal. But a firm will not appeal if the costs are high compared to the difference in fines (and bribes). So what is the optimal enforcement policy now?

It can be shown that for some parameter values it is impossible to implement the benchmark case. This in itself may not be very striking, because it may not be possible to implement the benchmark outcome even with honest inspectors if fines are constrained to be low (say, because wealth levels are low and there is limited liability constraint). But that is not the case here. In fact fines could be made large enough so that more deterrence than the benchmark case is achieved.

Suppose we turn our attention to situations where the benchmark action schedule can be implemented. Even then, this need not be the optimal policy if the monitoring costs are high. As mentioned earlier, the optimal policy in such situations might involve the legalization of some of the less harmful acts. Does the possibility of over-reporting make a difference to this result? For small values of the appeals cost, the presence of over-reporting possibilities does not make a difference to the legalization issue. But for higher values of appeals costs, implementing the first best policy requires greater enforcement cost, and hence the optimal policy would be to legalize the less harmful activity. The following proposition based on these arguments shows how optimal enforcement policy is diluted and distorted.

Proposition: Under-enforcement is more likely to be the optimal policy when inspectors are corrupt, information is soft, and cost of appeals or verification is high. Additionally, over-enforcement policy, which is never optimal under hard information, could turn out to be the optimal policy with soft information.

8.5 STEPS AGAINST CORRUPTION

We have argued that we need to pay more attention to the role of information in the study of corruption. The fact that information cannot be easily verified in many cases and also that corrupt officials are willing to manipulate such information makes the task of fighting corruption harder. We use two different examples in the context of a simple enforcement problem to show the implications of soft information.

Our first result, that some societies can get stuck in a low-compliance trap, can shed some light on the observation that similar societies could exhibit different levels of corruption. When a society is stuck in such a low-compliance trap, marginal changes in the incentive system (fight against corruption) are not likely to lead to significant improvements. One needs to introduce drastic measures to jolt the system out of the equilibrium. The second example shows how it becomes that much harder to achieve marginal deterrence when corrupt inspectors could manipulate information freely. Given that there are many instances where harmful activities differ in terms of severity (drugs, copyrights,

pollution, and bribery) and consequent social costs, our failure to achieve marginal deterrence entails significant welfare loss.

The key to both results is the feature that judicial costs are (relatively) high. This is an area where significant improvements could be made to dampen the effects of soft information. The spread of judiciary services, and a greater involvement of activist groups to enable less privileged groups better access to justice are needed. Additionally, improvements in technology (of information-gathering) and multiple sources of information could help in making information more easily verifiable.

REFERENCES

Becker, G. and G. Stigler (1974), 'Law Enforcement, Malfeasance and Compensation of Enforcers', *Journal of Legal Studies*, vol. 3, pp. 1–18.

Bierschbach, R.A. and A. Stein (2005), 'Overenforcement', *The Georgetown Law Journal*, vol. 93, pp. 1743–81.

Brunetti, A. and B. Weder (2003), 'A Free Press is Bad News for Corruption', *Journal of Public Economics*, 87(7), pp. 1801–24.

Chand, S.K., and K. Moene (1999), 'Controlling Fiscal Corruption', *World Development*, 27(7), pp. 1129–40.

Cotter, T.F. (2006), 'Fair Use and Copyright Overenforcement', *Legal Studies Research Paper* Series 6–69, University of Minnesota Law School.

Dasgupta, S. (2005), 'Rename REGA as Corruption Guarantee Scheme', *Pioneer*, 14 August.

Di Tella, R. and E. Schargrodsky (2003), 'The Role of Wages and Auditing During a Crackdown on Corruption in the City of Buenos Aires', *Journal of Law and Economics*, 46(1), pp. 269–92.

Dreze, J. (2008), 'India's Employment Guarantee Act: Reclaiming Policy Space', mimeo.

Fjeldstad, O.H. (2003), 'Fighting Fiscal Corruption: Lessons from the Tanzanian Revenue Authority', *Public Administration and Development*, 23(2), pp. 165–75.

Kofman, F. and J. Lawarree (1993), 'Collusion in Hierarchical Agency', *Econometrica*, vol. 61, pp. 629–56.

Klitgaard, R. (1988), *Controlling Corruption*. Berkeley and Los Angeles: University of California Press.

McMillan, J. and P. Zoido (2004), 'How to Subvert Democracy: Montesinos in Peru', *Journal of Economic Perspectives*, 18(4), pp. 69–92.

Mishra, A. (2005), 'Persistence of Corruption: Some Theoretical Perspectives', *World Development*.

—— (2006), 'Corruption, Hierarchies and Bureaucratic Structures', in S. Rose-Ackerman (ed.) *International Handbook on the Economics of Corruption*, Chetenham, UK: Edward Elgar.

—— (2007), 'A Model of Enforcement with Collusion and Extortion', mimeo, University of Bath, UK.

—— (2009), 'Corruption and Economic Growth', in *Perspectives on Corruption and Human Development* (UNDP), New Delhi: Macmillan.

Mookherjee, D. and I. P'ng (1994), 'Marginal Deterrence in Enforcement of Law', *Journal of Political Economy*, vol. 102, pp. 1039–66.

—— (1995), 'Corruptible Law Enforcers: How should They be Compensated?', *Economic Journal*, vol. 105, pp. 145–59.

Mookherjee, Dilip (1997), 'Incentive Reforms in Developing Country Bureaucracies: Lessons from Tax Administration', *Annual Bank Conference on Development Economics*, World Bank, DC.

Murphy, Kevin, Andrei Shliefer and Robert Vishny (1993), 'Why is Rent Seeking so Costly for Growth?', *American Economic Review*, 83(2), pp. 409–14.

Olken, Benjamin (2007), 'Monitoring Corruption: Evidence from a Field Experiment in Indonesia', *Journal of Political Economy*, 115(2), pp. 200–49.

Polinsky, A.M. and S. Shavell (2000), 'The Economic Theory of Public Enforcement of Law', *Journal of Economic Literature*, 38(1), pp. 45–76.

—— (2001), 'Corruption and Optimal Law Enforcement', *Journal of Public Economics*, vol. 81, pp. 1–24.

Rauch, J.E. and P.B. Evans (2000), 'Bureaucratic Structures and Bureaucratic Performances in LDCs', *Journal of Public Economics*, vol. 75, pp. 49–71.

Reinikka, R. and J. Svensson (2005), 'Fighting Corruption to Improve Schooling: Evidence from a Newspaper Campaign in Uganda', *The Journal of European Economic Association*, 2(2–3), pp. 259–67.

Rose-Ackerman, Susan (1978), *Corruption: A Study in Political Economy*, New York: Academic Press.

Sehgal, R. (2004), 'Triveni Devi and 109 other women vanquish corrupt ration shop dealers', *Infochange*. June, http://www.infochangeindia.org/features181.jsp (last accessed on 5 July 2009).

Shavell, S. (1992), 'A Note on Marginal Deterrence', *International Review of Law and Economics*, vol. 12, pp. 345–55.

Shleifer, A. and Robert Vishny (1993), 'Corruption', *Quarterly Journal of Economics*, vol. 108, pp. 599–617.

Taliercio, R., Jr. (2004), 'Administrative Reforms as Credible Commitment: the Impact of Autonomy on Revenue Authority Performance in Latin America', *World Development*, 32(2), pp. 213–32.

Tirole, J. (1986), 'Hierarchies and Bureaucracies: On the Role of Collusion in Organizations', *Journal of Law, Economics and Organizations*, vol. 2, pp. 181–214.

Triesman, D. (2000), 'The Causes of Corruption: a Cross National Study', *Journal of Public Economics*, vol. 76, pp. 399–457.

UNDP (2004), *Anti-corruption, Practice note*, New York: UNDP.

Van Rijckeghem, C. and B. Weder (2001), 'Bureaucratic Corruption and the Rate of Temptation: Do Wages in Civil Services Affect Corruption, and by How Much?', *Journal of Development Economics*, 65(2), pp. 307–31.

Wilde, L. (1992), 'Criminal Choice, Non-monetary Sanctions and Marginal Deterrence', *International Review of Law and Economics*, vol. 12, pp. 333–44.

World Bank (2007), *Strengthening WorldBank Group Engagement on Governance and Anti-Corruption*, Washington, DC: World Bank.

Yang, D. (2008), 'Integrity for Hire: An Analysis of a Widespread Customs Reform', *Journal of Law and Economics*, 51(1), pp. 25–57.

V

THE ROLE OF THE JUDICIARY IN FOSTERING EFFICIENT ENFORCEMENT OF LAW

9

ECONOMIC ANALYSIS OF PUBLIC INTEREST LITIGATION IN ENVIRONMENTAL CASES IN INDIA

Michael G. Faure and A.V. Raja

9.1 JUDICIAL ACTIVISM IN INDIA

Over the past few decades there has been a growing degree of what is known as judicial activism in India. The Supreme Court of India has become a policy maker in a whole range of isues such as human rights, environmental affairs, the functionings of political parties and municpal agencies, etc. This kind of activism also differs in character: interpretational activism, legislative activism, or exceutive activism.[1] Public interest litgation (PIL) is one such example wherein the Supreme Court of India (SCI) widened its scope and jurisdiction by interpreting the Consitution to expand the notion of fundamental rights and by relaxing the requirements of locus standi.

In particular, public interest litigation was used to complain to the courts about the inability of the executive, represented by the New Delhi municipal authorities to, control the air pollution levels in the capital. Studies have shown that the Supreme Court in India has played an instrumental role in reducing the increase of pollution levels, as far as ambient air quality in Delhi is concerned.[2] There is, in other words, convincing empirical evidence that public interest

[1] T.C.A. Anant and J. Singh (2002), 'Economic Analysis of Judicial Activism', *Economic and Political Weekly*, 26 October, pp. 4433–9.

[2] A.V. Raja and F. Xavier (2006) and Xavier and Raja (2008).

litigation in environmental cases is effective at least in some cases. The interesting question that remains is: what are the precise conditions that make public interest litigation in the specific case of India work? There is literature that analyses the separation of powers among the three organs of a constitutional democracy, the legislature, the executive, and the judiciary, using transaction cost arguments. Laffont and Meleu[3] have modelled the separation of powers as an instrument against corruption and have shown that the value of such separation is higher in developing countries.

Seen in this light, the phenomena of PIL is especially intriguing. The fact that a PIL was effective in the case of environmental harm raises pertinent questions about the quality of the other institutions, namely the executive and the legislature. At the same time there is a need to put this interesting phenomenon of public interest litigation in a broader law and economics framework to identify the specific conditions and circumstances under which public interest litigation can play a positive role in promoting the increase of social welfare in particular with respect to environmental problems. Such an analysis of course, needs to put public interest litigation in a broader perspective by comparing it to other environmental legal instruments, which may explain why under certain circumstances public interest litigation may have comparative benefits. This paper is a step in this direction.

The methodology to answer the above question will be both theoretical and empirical. At a theoretical level, the strengths and weaknesses of existing approaches to environmental problems in developing countries like India will be briefly reviewed. On the empirical side use will be made of the rich experience with public interest litigation as an issue developed by the Supreme Court of India. By transposing elements of the Indian case to a more general level we hope to critically examine environmental public interest litigation in India. For example, it is one thing to notice that pollution levels have come down as a result of the case law of the Supreme Court; it is still another to examine the costs at which this pollution abatement has been achieved, and whether public interest litigation is still an effective instrument or other mechanisms, such as public

[3] J. Laffont and M. Meleu (2001), 'Separation of Powers and Development', *Journal of Development Economics*, vol. 64, pp. 129–45.

enforcement or private litigation, would have achieved the same result in a less costly way.[4] Of course, the question also arises whether the results based on the specific case of India can be important for other cases as well and whether it is also possible at the theoretical level to predict the circumstances and conditions under which public interest litigation can be expected to play this positive role in environmental policy enforcement in other countries. It may be clear however, that by transposing elements of the Indian case to a more general level we will also critically examine the efficiency of public interest litigation in India.

Drawing upon the results of the earlier work,[5] where public interest litigation for the particular case of New Delhi has been analysed from an empirical perspective, the problem will now be reformulated by addressing the relevance for the general environmental law and economics doctrine. However, this is done within the following limits:

(1) We will not engage in actual modelling of the Supreme Court decisions, but rely on earlier work in that respect;
(2) We only address the public interest litigation in the field of environmental policy and not the rich case law in many other fields (such as human rights), and
(3) The goal of our paper is, of course, not to provide a final judgement on the efficiency of the decision-making by the Supreme Court in India, nor on its legitimacy. The latter is important since, notwithstanding the empirical evidence of success, there has also been some critical literature with respect to the judicial activism, more particularly from the perspective of accountability and legitimacy. Some have blamed the Supreme Court for engaging in 'judicial legislation'.[6] We merely aim at examining the phenomenon of public interest litigation within the traditional environmental law and economics framework.

[4] It has also been argued that specifying fuels and specific technologies is a poor second-best solution to notifying realistic emission levels and leaving the market to choose technology, engines and fuels. See Anant and Singh (2002).

[5] Xavier (2006) and Xavier and Raja (2008).

[6] S.P. Sathe (2001), 'Judicial Activism: The Indian Experience', *Washington University Journal of Law and Policy*, vol. 6, pp. 29–108.

The paper is organized as follows. First, in Section 9.2, we briefly summarise the traditional approaches, their weaknesses, and the traditional answers to environmental problems to explain the particular role of public interest litigation. In Section 9.3 we briefly address and define the phenomenon of public interest litigation more precisely. Section 9.4 provides a brief overview of how public interest litigation in India works in practice. Next, Section 9.5 addresses how the Indian Supreme Court has dealt with some of the practical problems that arise from a theoretical perspective. Having sketched the working of public interest litigation in India we provide an analysis and formulate a few key conditions for the role and significance of public interest litigation in an efficient environmental policy in Section 9.6. Section 9.7 is devoted to understanding the lessons from the Indian case. In Section 9.8 we summarise our conclusions: (1) A substantial reduction in pollution was seen in Delhi as a result of the intervention by the Supreme Court. (2) Earlier literature suggests that where the regulatory authorities are unable to be effective for reasons of apathy, corruption, etc. a second-best option is offered where those governing bodies less vulnerable to those reasons can step in. This has been seen in the case of public interest litigation in India.

9.2 TRADITIONAL APPROACHES TO ENVIRONMENTAL POLICY

The traditional approach in environmental policy is to first define, the ambient quality standards which determine the ideal pollution levels in a specific environmental component (air, soil or water).[7] These are referred to as ambient quality standards (in the US) or environmental quality standards (in Europe). The public interest model of environmental policy assumes that in the next step a cost-benefit analysis will be used to determine how specific standards should be set to arrive at the desired environmental quality. Once this is done, emission standards are laid down.[8] An important difference between an ambient quality standard and an emission limit value is that the ambient quality standard is often (although this can differ

[7] On optimal standard setting to control environmental pollution *see* Faure and Skogh, 188–207.

[8] Also referred to as emission limit values in Europe.

upon the legal system concerned) an ideal standard that should be reached in order to have an acceptable environmental quality. Ambient quality standards are therefore standards that address the regulator or licensing authorities, not individual polluters. It is the task of the licensing authorities to take these ambient quality standards into account when setting emission limit values for example, in specific permits of firms. If a firm were prosecuted for violating the ambient quality of the air or water surrounding their factory, then the problem would be to establish a causal link between the firm's pollution level and the deterioration of the ambient quality to prove liability. These problems of proving causation are avoided by focussing on environmental policy on emission limit values. Hence emission limit values are enforced upon individual polluters.

Once the next emission limit values are defined, they are then enforced upon polluters. The first-best, cheapest, and easiest solution is to use private law remedies like nuisance, tort, or in rare environmental cases, contract law. Shavell's classic paper[9] however, gives many reasons why private enforcement of these standards may not work: the damage may be too widespread as a result of which the individual damage for every victim is too low. They will therefore show, in the words of Schäfer, a 'rational disinterest'.[10] Therefore no individual victim would want to litigate. Moreover, especially in the environmental case, there can be causation problems and the time gaps between the emission and the actual occurrence of harm can be long (the so-called latency problem). As a result, a private law suit may never be brought.[11] In addition insolvency problems can arise; polluters can sometimes not be identified, etc. For all of these

[9] S. Shavell (1984), 'Liability for Harm versus Regulation of Safety', *Journal of Legal Studies*, vol. 13, pp. 357–74.

[10] H.B. Schäfer (2000), 'The Bundling of Similar Interest in Litigation, the Incentives for Class Action and Legal Actions Taken by Associations', *European Journal of Law and Economics*, p. 183.

[11] See generally on the influence of time on court cases (L. Kaplow (1986), 'An Economic Analysis of Legal Transition', *Harvard Law Review*, vol. 99, pp. 509–617) and for the consequences on insurability (M. Faure and P. Fenn (1999), 'Retroactive Liability and the Insurability of Long-tail Risks', *International Review of Law and Economics*, pp. 487–500). Latency is traditionally advanced as an argument in favour of regulation. See Shavell (1984).

reasons there is, in principle, a strong case for public enforcement of environmental standards.

A third step in the literature is that especially as far as developing countries are concerned, many weaknesses appear in the traditional public interest explanation as it was just presented. Failures can occur at the standard-setting level, more particularly through the influence of private interest, as a result of which regulatory solutions will be suboptimal. Moreover, many failures can also occur at the level of enforcement. Sometimes these failures are rather innocent (e.g., due to the fact that enforcement agencies lack capacity or instruments to act against polluters); in other cases a collusive relationship between regulators and environmental polluters inhibits an effective enforcement of environmental standards. Ample evidence can be seen (not just in developing countries) of these collusive practices which result in under-enforcement. Additionally, like in many developing countries, when there are outright corruption problems, public enforcement will obviously fail to be effective.[12]

A fourth step has already been taken in the law and economics literature. The traditional answer to weaknesses in certain standard-setting mechanisms or enforcement tools has been to use (a combination of) instruments that are less vulnerable to these problems. An impressive overview of the possible remedies and their combination can be found in the well-known work of Gunningham and Grabosky.[13] For example Shavell[14] argued that since regulation may be outdated fast and cannot cover every possible situation, and since there may never be 100 per cent enforcement, private enforcement through tort rules should supplement regulation. Hence, tort rules would still have a complementary function to back up weaknesses in regulation. Also Schäfer argued that in systems where enforcement capacity is lacking there may be less reliance on vague standards (which require administrative capacity for implementation and enforcement) and a

[12] A. Ogus (2004), 'Corruption and Regulatory Structures', *Law and Policy*, vol. 26, pp. 329–46 and (2007) 'Regulatory Arrangements and Incentives for Opportunistic Behaviour', in T. Eger, M. Faure, and N. Zhang (eds), *Economic Analysis of Law in China*, Cheltenham, Edward Elgar, pp. 151–63.

[13] N. Gunningham and P. Grabosky (1998), *Smart Regulation. Designing Environmental Policy*, Oxford: Clarendon Press, 1998.

[14] Shavell (1984).

greater use of specific rules.[15] Also Ogus[16] has powerfully argued that when enforcement may be inefficient, given corruption problems, one should move to policy instruments that may be less vulnerable to collusive practices: for example use the court system rather than public enforcement mechanisms that are more vulnerable to corruption.

To some extent these suggestions and remedies are all very useful but may not always provide a very practicable solution in the specific case of India. For example, from earlier studies[17] it appeared that an alternative to private enforcement by victims as a remedy for collusive practices at enforcement level was often not possible in cases where the harm remained very widespread. As a result, an individual victim lacked an incentive to file a lawsuit since this could only lead to recovery of a relatively small damage award. Equally they indicated that class actions, although theoretically possible,[18] often were not initiated in the specific case of India, especially because the transaction costs for the group to get organized were simply too high. There is in addition a third reason why a private legal remedy may not always work: Assume there is a polluter who is easily identifiable but whose pollution causes harm in the vicinity of the industry. The pollution will be felt most intensely by those who live in and around the factory and less intensely by those who live progressively further away. The damage suffered by the victims is therefore not uniform, but the totality of the damage can be large. Since not all the victims of the pollution have uniform power, nor incentives to litigate (also depending upon different income levels), not all may file a lawsuit and a class action may not emerge. A potential development is that the polluter would try to identify the victim types who have the highest propensity to litigate (most activist, high damage, and high income, and thus the possibility of filing a law suit) and would then try to negotiate a settlement with them. The polluter thus can prevent problems by concluding a settlement only with the potential litigators.

[15] H.B. Schäfer (2006), 'Rules versus Standards in Rich and Poor Countries: Precise Legal Norms as Substitutes for Human Capital in Low-income Countries', *Supreme Court Economic Review*, vol. 14, pp. 113–35.

[16] Ogus (2004, 2007).

[17] A.V. Raja and F. Xavier (2007), *Strategic Behaviour with Differential Damages*, Working Paper, Department of Economics, University of Hyderabad.

[18] Schäfer (2000).

A problem from the social welfare perspective is that through this partial bribe the probability of litigation is seriously reduced, but that the externality may only be partially internalized.

This example shows the potential of public interest litigation, where the potential litigators could become an impossibly large set of people. At the same time the identification ex ante of these litigators would become too costly for the polluter. Hence in this way opportunistic behaviour by the polluter (which would be the result under the application of traditional tort rules) could be avoided.

Before addressing whether public interest litigation can thus potentially be an effective solution for the above-mentioned problems we first have to identify what is exactly meant by public interest litigation.

9.3 WHAT IS PUBLIC INTEREST LITIGATION?

Public interest litigation in India was initiated by a few Judges of the Supreme Court, as a method to redress public grievances. It was meant as a protective mechanism to redress the basic violation of human rights of the poor and the needy.[19] It was also a way to address concerns about government conduct or policy if that was seen to be against the interests of society. More specifically, public interest litigation was viewed as solving the problem of 'access to justice' in developing countries where, large sections of the population are neither literate enough nor do they have the means to use the legal system to redress the problems that may have arisen due to violation of fundamental rights. Frequently, the poor and the underprivileged are unwilling to assert their rights because of poverty, ignorance or, the fear of social or economic reprisals from the dominant sections of the society. These disabilities could be reduced if the law allows a concerned citizen to sue in a court of law. The classical theory of requiring a locus standi precludes such a 'representation',[20] and thereby keeps the grievances of the poor from reaching the courts.

The report to the Government of India, Ministry of Law, Justice and Company Affairs, by Bhagawati, J. and Krishna Iyer, J.,[21] expressly

[19] For an evolution see Sathe (2001).

[20] In many such cases, the affected may not even know the identity of the concerned citizen.

[21] *Report on National Juridicare: Equal Justice-Social Justice, 61(1977).*

recommended the lowering of the locus standi as a means of allowing concerned citizens to file cases on behalf of the underprivileged.[22] Thus the meaning 'persons aggrieved' was widened. The second modification to the standing doctrine was that a concerned citizen (or a voluntary organization) could sue, not just as a representative of others but in his or her *own right* as a member of the citizenry to whom a public duty is owned; this may be termed 'citizen standing'. Although this relaxation was primarily justified on the grounds of the 'access to justice' principle, it was increasingly used to check the abuse of the executive authority in a modern welfare state. In India regulation was the primary method to controlling environmental harm. This gave the government's regulatory competences enormous power. On occasion, a misuse of power and authority did happen. Such misuse may lead to injury that may be diffused, for example, where government policy or inaction may threaten the environment. In such cases traditional doctrines of standing would preclude citizens from seeking protection. The Supreme Court expanded the standing to enable citizens to challenge such government actions in the public interest even though the citizen himself suffered little or no harm.[23]

9.4 PUBLIC INTEREST LITIGATION IN INDIA

In India the citizen has three civil remedies that s/he may choose from to obtain redress for environmental damages: (a) a Common Law tort action against the polluter, (b) a writ petition to compel the agency to enforce the law and to recover damages from the violator, or (c) an application for compensation under the National Environmental Tribunal Act of 1995 or the Public Liability Insurance Act of 1991.

However, sub-optimal damage compensations and the lengthy adjudication of cases, combined with chronic inflation dilute the value of damages and offer no incentives for private litigation.[24] Within this

[22] 'On behalf' does not mean that it was in consultation with or with the consent of the affected people. Any concerned citizen who felt that an injustice was being perpetuated on large sections of society could move the courts.

[23] In the *Ganga Pollution (municipalities)* Case (*Mehta* v. *Union of India*, AIR 1988 SC1115), the Supreme Court upheld the standing of a Delhi citizen (the well known environmental activist and lawyer Mr Mehta) to sue the government agencies whose prolonged neglect had resulted in severe pollution of the river.

[24] S. Diwan and A. Rosencranz (2001).

form of redress, injunctions rather than damages seem to be more prevalent. However, the cases apply when the source of pollution and the victims are well identified and the plaintiffs have locus stand in court. Close to the public interest litigation model that evolved is the model of public nuisance. Since a public nuisance interferes with a public right, it is not tied to the violation of private property rights, and every *affected* citizen can make use of this remedy. Commenting on one such case, Krishna Iyer, J. observed, 'At issue is that branch of Public Law bearing on the community actions and the court's power to force public bodies under public duties to implement specific plans in response to public grievances.'[25]

The writ jurisdiction that empowers the Supreme Court of India, under Articles 32 and 226 of the Constitution, to issue directions orders or writs of *habeas corpus, mandamus, prohibition, quo warranto*, and *certiorari*. There have been cases where a writ of mandamus has been successfully used to command action by a public authority when it was found that the authority failed to carry out its duties.[26]

Public Interest Litigation is an extension of writ jurisdiction. However, there are a few differences: (a) The court itself can initiate action in public interest if the judge concerned comes to know of any violations and (b) The locus standi is relaxed so that any public spirited citizen can move the court.

The following are some of the salient characteristics of the Public Interest Litigation process in India:

1. Public Interest Litigation has recently adopted the notion of a continuous mandamus. This involves securing compliance through periodic 'interim' directions. Here, the court does not formally admit the petition but prods the recalcitrant agency to perform its duty within a time frame. The case is posted every

[25] Krishna Iyer, in Municipal Council, *Ratlam* v. *Vardhichand*, AIR 1980 SC 1622, in S. Diwan and A. Rosencranz (2001).

[26] *Rampal* v. *State of Rajasthan*, AIR 1981 RAJ 121. The residents of a town filed a petition showing that the Municipal authorities failed to discharge their duties in the removal of stagnant waters and garbage. The court directed that Municipal authorities not only clean up but also construct the appropriate drainage or sewers for the discharge of such water.

few months to monitor the agency's performance and for further interim directives.

2. Parties and official agencies may be joined (and even substituted) as the litigation unfolds: and new and unexpected issues may emerge to dominate the case.
3. The expanded status includes at least two different kinds of standings recognized by the Supreme Court: (i) 'Representative Standing' to represent the poor or the underprivileged. (ii) 'Citizen Standing' in cases of executive inaction or abuse.[27]
4. The dismissal of Public Interest litigation if found to be motivated by private interest usually imposes costs on the litigant. In one case the court explicitly noted that:

> Personal interest cannot be forced through the process of this court under Art. 32 of the Constitution in the garb of a Public Interest Litigation ... It is the duty of this court to discourage such petition and to ensure that the course of justice is not obstructed or polluted by unscrupulous litigants by invoking the extraordinary jurisdiction of this court for personal matters under the garb of the public interest litigation. (Supreme Court of India in the case of *Subhash Kumar vs State of Bihar*, AIR 1991, SC 420, quoted in Diwan S. and A. Rosencranz (2001)).

5. The Court has the inherent power to appoint independent commissions to verify the facts of the case. In the case of environmental problems, the National Environment Engineering Research Institute (NEERI) is called upon frequently by the courts to submit reports.
6. The judicial 'notice of facts' reduces the litigant's evidentiary burden. In the Ganga pollution case the court did not wait for hard scientific proof on the health effects of the population and the damage to property. In view of certains books, and writings, the court assumed that such injuries had either occurred or are likely to occur and proceeded to issue remedial directions.
7. In recent PILs the 'reversal of the burden of proof' was used by the courts as a 'precautionary principle'. In that case the onus of

[27] Early Environmental cases decided by the SCI (Supreme Court of India) resulted in the closure of lime stone quarries in the Dehradun region, the installation of safeguards at a chlorine plant in Delhi, the closure of polluting Tanneries on the Ganga are examples of citizen standing cases. See S. Diwan and A. Rosencranz (2001).

proof is on the developer or the industry to show that his or her action is environmentally benign.

Some of the weaknesses of the public interest litigation India are the following:

1. The most fundamental issue that is raised by public interest litigation as a process is that in the name of upholding fundamental rights and the protection of individual liberty, the courts frequently overstep their jurisprudence and involve themselves in matters that belong to the executive and the legislature. This, however, has not been overlooked by courts. In a judgment in the case of *Bandhua Mukti Morcha* v. *Union of India* Pathak, J observed:

   In the process of correcting executive error or removing legislative omission the court can so easily find itself involved in policy-making of a quality and a degree characteristic of political authority, and indeed run the risk of being mistaken for one. An excessively political role identifiable with political governance betrays the court into functions alien to its character and tends to destroy the delicate balance envisaged in our constitutional system between its three basic institutions... It is a serious question whether in every case the same awesome respect and reverence will endure during different stages of affirmative action seeking to regulate the lives of large numbers of people, some of whom never participated in the judicial process.

2. There is very little by way of theory that would explain why a public spirited citizen would move the courts. Certainly this is an important question from a law and economics perspective. If Rational Choice theory is to be believed, then we must conclude that every public interest case also generates some private gains apart from those pertaining to the case itself. In some cases, the prime mover of the litigation gains favour with the courts as an expert on such matters and may help the courts in future cases. This in itself may not be a bad thing, but it could lead to problems: A much discussed case is that of M.C. Mehta whose fame as the prime mover of the PIL in the Delhi Air pollution case and its subsequent success led him to file a petition to include 'environmental Studies' as a compulsory subject in the Under-graduate teaching curricula of all schools and colleges. The Supreme Court, instead of putting forward the suggestion

to the concerned Ministry of Education, proceeded to direct the University Grants Commission to implement the suggestion without any further delay. Was the S.C.I justified in doing this? The second danger is that PIL cases may be filed for issues that give better media coverage rather than on issues that are of fundamental importance but which do not have dramatic developments. A case in point was that of a PIL that was filed against a beauty pageant held in Bengaluru on the grounds that it was against the morality of Indian women.

3. That the relaxation of the standing requirement would result in a mushrooming of PILs has been borne out in the case of India.[28]

Having sketched an overview of how public interest litigation in India works, we analyse in the next section how the Indian judiciary deals with Public Interest Litigations in more detail.

9.5 THE SUPREME COURT OF INDIA AND PUBLIC INTEREST LITIGATION

9.5.1 Plaintiffs

We already mentioned that the requirement of standing has been relaxed in a considerable way generally by the Supreme Court in public interest cases. Hence, the eligibility of a person to invoke the jurisdiction of the courts is very much enlarged, and a person can simply seize the court by writing a letter without having to pay for a lawyer.[29] Rajamani[30] states that in specific environmental cases, public-spirited individuals have seized the court inter alia to seek the protection of the Taj Mahal from corrosive air pollution,[31] to reduce discharges into the Ganga,[32] or to address the air pollution in Delhi and other metropolitan cities.[33]

[28] See Raja and Xavier 2008.

[29] Sathe (2008), pp. 74–75.

[30] L. Rajamani (2007), 'Public Interest Environmental Litigation in India: Exploring Issues of Access, Participation, Equity, Effectiveness and Sustainability', *Journal of Environmental Law*, vol. 19 (3), p. 295.

[31] *M.C. Mehta* v. *Union of India* (Taj Trapezium Case), Writ petition number 13381 of 1984.

[32] *M.C. Mehta* v. *Union of India* (Ganga Pollution Case), Writ petition number 3727 of 1985.

[33] See inter alia *M.C. Mehta* v. *Union of India* (Delhi Vehicular Pollution Case), Writ petition number 13029 of 1985.

Apparently, the individuals who wish to seek the Supreme Court in the public interest to address a particular environmental concern do not encounter any difficulty as far as the formal locus standi requirements are concerned. Given the high number of cases that have effectively been brought (even though it is usually a small group of individuals that bring the cases), one can argue that their altruistic motives have apparently provided them sufficient incentives to bring the mentioned cases in the public interest.

However, there is a growing concern, given the increasing number of cases before the Indian Supreme Court,[34] that PIL suits would be filed for reasons that would be vexatious or frivolous. As a result of this, some have raised the question whether guidelines should be set to restrain and limit public interest litigation.[35] However the judges are not unaware of the problem:

> Personal interest cannot be enforced through the process of this court under Art. 32 of the Constitution in the garb of public interest litigation ... It is the duty of this court to discourage such petition and to ensure that the course of justice is not obstructed or polluted by unscrupulous litigants by invoking the extraordinary jurisdiction of this court for personal matters under the garb of the public interest litigation.[36]

9.5.2 Defendants

A particular aspect of public interest litigation in India is that usually litigants proceed against the state, asking the state to take a particular action where regulation has failed or lacks enforcement. This is generally the case, but particularly so, as far as the environment is concerned. In most cases it is effectively the state who is the defendant and who is expected to help the court to find the truth rather than fight the allegations. In that sense in public interest litigation there is no traditional adversarial proceeding.[37]

This may of course have consequences, more particularly in environmental cases, to the extent that plaintiffs require that the

[34] According to Krishnan (2008) in March 2007 there were 40,000 cases pending before the Supreme Court of India.

[35] Rajamani (2007) p. 321.

[36] Supreme Court of India in the case of *Subhash Kumar* v. *State of Bihar*, AIR 1991 SC 420, quoted in Diwan and Rosencraz (2001).

[37] Sathe (2001), p. 77.

court should order the state to undertake a particular action which may have consequences for third parties. These third parties are of course usually the industries involved who will have to take action as a result of the court's order. Remarkably, these industries that are affected by the court's order usually do not appear as defendants in the case. Two remarkable cases discussed by Rajamani[38] can illustrate this. One major case initiated by litigator M.C. Mehta dealt with the dramatic consequences of air pollution in Delhi for human health. After a long procedure and consulting (to which we will refer below) the court ordered inter alia the conversion of the entire city's bus fleet to compressed natural gas (CNG) as an alternative vehicular fuel on the basis that it was less polluting than petrol or diesel.[39] Even though some consultation had taken place, the private bus operators (who were of course largely affected by the case) were not consulted by the Environment Pollution Prevention and Control Authority (EPCA) which was authorized by the judiciary. Private bus operators felt frustrated by their lack of access to the EPCA and to the court.[40] Moreover, since the private bus operators[41] did not formally participate in the proceedings before the court they were only notified by the Delhi government fifteen months after the court's order. The lack of involvement of these stakeholders that have to carry out the court's order may therefore lead to problems as far as information on efficient pollution reduction measures is concerned and at the level of enforcement.

A similar problem arose in a municipal solid waste management case where, as a result of a court order, the Ministry of the Environment and Forestry (MoEF) notified, municipalities about so called municipal solid waste management and handling rules (so called MSW rules). They had to be implemented by municipalities across India, but these, not being involved in the procedure, apparently were not able to meet the deadlines set by these MSW rules.[42]

[38] Rajamani (2007).

[39] Ibid., pp. 298–9.

[40] Ibid., p. 305.

[41] Who accounted for almost 80 per cent of the public transport in Delhi, according to Rajamani (2007), p. 299.

[42] Rajamani (2007), p. 311.

9.5.3 THE COURT

9.5.3.1 Information Advantage

The first question is how the Supreme Court of India is able to get that actively involved in environmental policy at all, given the fact that traditional law and economics literature has in that respect pointed at the informational advantage of administrative agencies. To a large extent, this will depend upon what the court is actually doing. If the court were merely to compel the government to enforce its own legislation, information problems should not necessarily arise—since the environmental standards would have been set already by the executive and the court would merely enforce a prior decision of the executive. Indeed, to some extent environmental cases are directed against the government and have as their goal an intent to compel the government to do its duty, or to prevent the government from doing what it was legally forbidden to do.[43]

For example in the Delhi vehicular pollution case Mr Mehta started a case arguing that already existing environmental laws obliged the government to take steps to reduce Delhi's air pollution in the interest of public health.[44] However, the mere fact that the court would 'only' compel the government to do something which it was already forced to do by law does not mean that no information problem arises at all. For example, even in the case of the Delhi vehicular pollution, the question still arises as to *how* the government should order polluters to reduce air pollution. The court therefore goes well beyond the mere contents of a positive duty to act by examining *inter alia* technologies worldwide, and available technologies in India so as to finally order the conversion of the entire bus fleet in Delhi to CNG.[45] Moreover, in many cases the Supreme Court goes well beyond the mere execution of a prior existing legislation. An important basis for the judicial activism by the Indian Supreme Court can also be found in the possibility of judicial review provided for in Article 13 (1) of the Indian Constitution. Article 13 (2), moreover, provides that states shall not make any laws that take away or abridge any of the fundamental rights, and that laws made in contravention shall be void.[46]

[43] Sathe (2001), p. 80.
[44] Rajamani (2007), p. 298.
[45] Ibid., p. 298–9.
[46] Sathe (2001), pp. 38–9.

The court has a variety of ways to deal with the information problem, more particularly in the environmental cases we already mentioned. For example in the municipal solid waste management case the court set up a committee involving both the Ministry of Urban Development (MoUD) and the Ministry of Environment and Forest (MoEF). Hence, the court can rely on the available expertise within these ministries.[47] Also in the Delhi vehicular pollution case, the court inter alia directed the MoEF to set up a committee chaired by Saskia, J to assess technologies worldwide, technologies available in India and recommend 'low cost alternatives for operating vehicles and reducing pollution levels in the metropolitan cities of India'.[48] The fact that in both cases the court created committees shows that the court relies on the input of stakeholders in the committees to come to its decision. For example, in the vehicular pollution case the committee set up by the court contained high ranking bureaucrats, an NGO, and a representative of the automobile industry.[49]

To some extent the working of these committees could thus inform the court of appropriate measures to reduce pollution effectively. Moreover, these decisions by the Supreme Court were never taken on a very short term. In some cases many orders have been issued by the court and the Supreme Court has hence been following the cases on a long term, adapting its orders to increasing insights based on information provided by the committees.

9.5.3.2 Legitimacy

The strong involvement of the court over a long term leads to an effective replacement of environmental policy-making by the executive to an environmental policy-making by the judiciary. This, of course, raises questions of the legitimacy and accountability of the court to fulfil this role. Sathe notes that the Indian Supreme Court not only makes law in the sense that some view the role of case law in interpreting statutes, but 'actually legislates'. The court thereby 'transcended the limits of the judicial function and has undertaken functions which really belong to either the legislator or the executive. Its decisions

[47] Rajamani (2007), p. 297.
[48] Ibid., p. 298.
[49] Ibid., p. 303.

clearly violate the limits imposed by the doctrine of separation of powers'.[50] This role of the court is criticized for the simple reason that the court lacks the institutional equipment for undertaking legislative or administrative functions[51] and is considered a problem because, judges are not elected, they are not accountable to any constituency.[52]

Anant and Mishra[53] have modelled the costs and benefits of activism. Their conclusion is noteworthy:

> To the extent benevolent individuals of an institution choose to be activists, activism can be beneficial as it acts as a check against corrupt practices by non-benevolent individuals in other institutions. Such benefits are likely to be higher in countries where enforcement initiatives are weak (transaction costs associated with bribery are low) and some institutions exhibit high levels of non-benevolence.

This is also clear in the environmental cases where the court establishes itself as the main protector of the environment[54] and undertakes tasks that go far beyond the judicial function. For example in the Delhi vehicular pollution case the court not only ordered the entire city bus fleet to convert to CNG but also got involved in the issue of CNG pricing.[55] Thus, in many cases where formal legislation is lacking, the court has no problem in making laws for example, by defining guidelines to prevent sexual harassment in the work place.[56]

The examples show that in practice the Supreme Court goes indeed far beyond the traditional competences of a judiciary, violating the separation of powers which raises fundamental questions of legitimacy and accountability.

9.5.3.3 Independence

How is it possible that in a developing country like India with all its capacity and corruption problems at the level of the executive, the judiciary has apparently gained such a moral status that it is

[50] Sathe (2001), p. 88.

[51] Ibid., p. 89.

[52] Ibid., p. 105.

[53] A. Mishra and T.C.A. Anant (2006) 'Activism, Separation of Powers and Development', *Journal of Development Economics*, vol. 81, pp. 457–77.

[54] Rajamani (2007), p. 318.

[55] Ibid., p. 313.

[56] Ibid., pp. 316–17.

effectively able to act where politicians and administrators fail? The traditional answer in law and economics scholarship is that this can only happen in an environment where guarantees are created for the court's independence, thus preventing judges from being subject to the same corruption problems as politicians or bureaucrats.[57] It would go beyond the scope of this paper to examine why it has not been possible to realize in India what has traditionally been a major problem in other south-east Asian developing countries (like Indonesia), where it is believed that the Supreme Court of the land is captured by the political party in power. We can suffice by referring to the cases we already mentioned which apparently show that the justices of the Supreme Court have no difficulty of acting de facto independently by ordering the government for example to take measures to reduce air pollution in Delhi, or to reduce the discharge of polluting substances into the Ganga.

Sathe[58] provides a few historical reasons why the judiciary in India has, since the early years of the independence, developed itself as a defender of the young democracy. Some guarantees suggested by law and economics scholarship are apparently incorporated into the Indian model. In this respect we can, for example, point at the fact that a judge is removable only on very specific grounds (misbehaviour or incapacity), but that this must then be assessed by a quasi-judicial body. Politicians can hence not remove a judge unless a charge of misbehaviour is found valid by the committee. Also, the privileges or allowances, including the pension rights, of the judge cannot be varied to his/her disadvantage after his/her appointment.[59]

[57] In general however, corruption in the Indian judiciary is a well documented fact: According to a survey by media house, 80 per cent of the population sample agree that there is rampant corruption in the lower courts.

[58] Sathe (2001), pp. 99–100.

[59] The independence of the judiciary is guaranteed by the Constitution of India which enacts that every Judge of the Supreme Court will hold office until, he attains the age of 65 years and that every Judge of the High Court until the age of 62 years. The Parliament is authorized to prescribe the privileges, allowance, leave, and pension of the Judges of the Supreme Court, subject to the safeguard that these cannot be varied during the course of tenure of the judges to their disadvantage. No Judge shall be removed from his/her office by the President except upon the presentation of an address by each House of Parliament for such removal on the grounds of proved misbehaviour or incapacity. A Judge of

In a remarkable recent study, Hayo and Voigt[60] explain de facto judicial independence. They develop a few parameters on the basis of which they examine both *de juri* and de facto independence. Not surprisingly, India scores well in that empirical test.

9.5.3.4 Court versus Executive

A question which is equally important is how it is possible that the Supreme Court is able to do what it is doing and why, also in environmental matters, politicians and bureaucrats apparently comply with the orders of the court. Some of this of course has to do with the moral status of the court resulting from its independence; another aspect is related to the instruments available to enforce its orders, which we will elaborate in the following sections (see 9.5.5), but there are some other reasons why politicians generally accept the judicial activism by the court. Sathe holds that, strikingly, none of the political players has strongly protested against judicial intrusion into matters that essentially belong to the executive. One reason may be that political players hold that in matters involving conflict between various competing interests, the courts are better arbiters than politicians.[61] Thus, a traditional public choice explanation could be provided for the success of the court: politicians may simply fear to lose votes either when favouring environmental interests too strongly to trade unions or representatives of industry, or the reverse could happen if they favoured industrial interests too clearly. Notwithstanding the fact that politicians do not generally criticize judicial activism, it has become a subject of controversy in India and some attempts have been undertaken to hinder the power of the

---

the Supreme Court or the High Courts is appointed by the President of India in consultation with the Chief Justice of India and such of the Judges of the Supreme Court and the High Courts as s/he may deem necessary. The Constitution of India insulates the Supreme Court and the High Courts from political criticism, and thus ensures their independence from political pressures and influence by laying down that neither in Parliament nor in a State Legislature can the conduct of a Supreme Court or High Court Judge in the discharge of his duties be discussed. (Confederation of Indian Bar International Seminar on 'Judicial Independence' 23 November, 2007.)

[60] B. Hayo and S. Voigt (2007), 'Explaining *De Facto* Judicial Independence', *International Review of Law and Economics*, vol. 27, pp. 269–90.

[61] Sathe (2001), p. 89.

courts.[62] But apparently none of these attempts has so far been able to stop judicial activism resulting from public interest litigation in the environmental field.

9.5.4 What Does the Supreme Court Order?

The specific environmental cases also make clear what it is that the Supreme Court actually orders. These orders are usually directed against the government or public authorities. The Constitution gives the Supreme Court and the High Courts the power to use 'directions, orders, or writs' for achieving the objectives of those articles. Thus, the court can issue specific orders to parties to take or refrain from specific action.[63] For example, in *M.C. Mehta* v. *India*, the court asked the Government of India to explain why a large part of the toll tax and the visitor's fees received from tourists visiting the Taj Mahal should go to the Agra Development Authority instead of being spent on the preservation of the Taj Mahal and the cleaning of the city of Agra.[64] We already mentioned that in the Delhi vehicular pollution case the court ordered in July 1998 inter alia the conversion of the entire city bus fleet to CNG by the 31 March 2001.[65]

9.5.5 Enforcement

Of course, a crucial question is how one can guarantee compliance with the orders of the court. The formal answer is that Article 141 of the Indian Constitution provides that the law declared by the Supreme Court shall be the law of the land. The High Courts have held that this is not only the case for the *ratio decidendi*, but even for the *obiter dicta* of the Supreme Court. With that, the orders of the Supreme Court are therefore also the law of the land.[66] That of course does not immediately answer the question why defendants (or even third parties) comply. The most important reason is apparently that in India there is a strong fear of being held in contempt by the court, a risk that not only the defendant runs, but all those who are affected

[62] Ibid., p. 30.
[63] Sathe (2001), p. 84.
[64] Ibid., p. 83.
[65] Rajamani (2007), p. 299.
[66] Sathe (2001), p. 86.

by the court's order. For example in the Delhi vehicular pollution case, the Delhi buses that could not comply with the court's order (to convert to CNG) simply stayed off the roads for fear of being held in contempt.[67] Also, in the municipal waste case, municipalities across India reacted panic-stricken on being asked to implement the court's orders, induced by the threat of contempt proceedings.[68] Of course, the court tries to facilitate compliance with its orders, for example, by providing an implementation schedule. More particularly in the municipal waste case, the court ordered on 3 October 2000, the compliance with the new (so called MSW) rules which required the setting-up of waste processing and disposal facilities by 31 December 2003, the improvement of existing landfill sites by 31 December 2001, and the identification of landfill sites for future use by 31 December 2002.[69] Notwithstanding these timelines and schedules, municipalities had great difficulties meeting the court-imposed deadlines.[70]

In the Delhi vehicular pollution case, the court moreover used the instrument of penalty payment to enforce its decision: it ordered private diesel buses to pay fines of Rs 500 per day in case of non-conversion to CNG.[71]

Notwithstanding the fact that the literature reports that more particularly the affected third parties (like the private bus companies or municipalities) may encounter difficulties in complying with the court's order, there is almost never a discussion on whether there should be compliance at all. Apparently the threat of being held in contempt, and (when it is has been ordered) the fear of having to pay the penalty are sufficient deterrents to induce compliance with the court's order.

9.6 EFFICIENCY AND/OR EFFECTIVENESS OF ENVIRONMENTAL PUBLIC INTEREST LITIGATION

9.6.1 Efficiency

It may be clear that even if one were to find some evidence that public interest litigation in India of the type described in this paper were to

[67] Rajamani (2007), p. 300.
[68] Ibid., p. 311.
[69] Ibid., p. 297.
[70] Rajamani (2007), pp. 311–12.
[71] Ibid., p. 301.

reduce pollution levels, it does not at all mean that pollution levels would necessarily be reduced to efficient levels. Indeed, we did not make any analysis of what efficient pollution levels in India would be and whether the public interest litigation is the ideal instrument to reach that goal. It is probably also impossible to test in practice the efficiency of public interest litigation in the sense that one could test whether this is the legal instrument that optimally contributes to maximizing social welfare. What one can at best test is whether the type of public interest litigation we described in this paper has had any environmental effectiveness in the sense that pollution levels were actually reduced. But even if one were to find that there is such an environmental effectiveness, this does not say too much from an economic perspective for the simple reason that the price to reach a reduction of environmental pollution may have been much too high compared to alternative solutions. Hence, in addition to addressing the environmental effectiveness of public interest litigation, one should also pay some attention to its cost-effectiveness by addressing whether the goal set by the legislator (or court) of reducing pollution levels has, in this particular case, been reached at the lowest possible costs.

9.6.2 (Environmental) Effectiveness

A question that could be asked is whether the public interest litigation and the actions of the Indian Supreme Court in the environmental arena have effectively reduced pollution levels in the cases where the court intervened. The evidence in that respect is not overwhelming. On the one hand, the possibility of a court to intervene in pollution cases is of course inherently limited by the fact that the court only acts in individual cases and does not have the same authority as a legislator or administrative authority to regulate pollution levels more generally. Sathe therefore holds that the court 'cannot stop entirely the degradation of the environment or government lawlessness. Its actions in these areas are bound to be symbolic'.[72] Nevertheless, there is some empirical evidence that the orders by the Indian Supreme Court have led to a substantial reduction of pollution levels. The starting point for our paper were some of the papers written by

[72] Sathe (2001), p. 89.

Raja and Xavier[73] who argued that inter alia the already-mentioned decision of the Supreme Court in the Delhi vehicular pollution case has led to substantial improvements of the air quality in Delhi. Also Rajamani[74] holds that the decisions of the Supreme Court in the municipal solid waste management and Delhi vehicular pollution cases have led to visible improvements in solid waste management in cities and in the air quality in Delhi. Nevertheless, the actions of the Supreme Court in this respect do not go undisputed. We have already mentioned that a lack of involvement of stakeholders (more particularly the defendants) led not only to frustration (for example, by the private bus operators and municipalities) but also to delays in the implementation of the court's decisions. Rajamani even argues that the court's decision to force Delhi bus companies to convert to CNG initially led to a chaotic situation whereby buses had to travel 40 kilometres and spend several hours queuing up outside CNG dispensing stations (given too few dispensing stations were available), leading to too few buses being available.[75]

9.6.3 Cost-effectiveness

From an economic perspective it is especially important to ask the question whether the public interest litigation was cost-effective in the sense that the environmental results (of reducing pollution levels) were also reached at the lowest costs possible. This question has two separate aspects. On the one hand the question arises whether the measures imposed as a result of the Supreme Court ruling were cost-effective in the sense that the contents of the environmental measures ordered by the court gave incentives to operators to reach the reduction of pollution levels at the lowest cost possible. On the other hand, one could equally ask whether public interest litigation is the lowest cost alternative to reach this particular goal. To start with this last question, information on the precise costs of public interest litigation is not known. One could argue that these costs could be

[73] Raja and Xavier (2006, 2008)

[74] Rajamani (2007), p. 319.

[75] Which constituted especially an equity problem since it was particularly the people who were relying on public transport and who thus initially suffered from the decision (see Rajamani, 2007, p. 308).

substantial, given that the Supreme Court has in some cases ordered several rulings, involved many committees, and had to follow-up a case over many years. The administrative costs for the functioning of the judiciary can thus be substantial. However, the essence of private interest litigation is precisely that a plaintiff can start the proceedings at relatively low costs (writing of a letter) and that, moreover, these (administrative) costs of public interest litigation should therefore not necessarily be larger than the costs of the functioning of a regulatory system. In the latter case an administrative agency needs to intervene to set regulatory standards for the entire industry, needs to install a monitoring, enforcement, and compliance system, etc. Hence, even if precise data fail, one can intuitively argue that the administrative costs of public interest litigation should not necessarily be higher than the costs of reaching the same goals via the regulatory system.

There is, however, some criticism on the fact that public interest litigations are often filed by the same type of petitioners, what some would characterize as a 'clique'. The danger of this (often the same plaintiffs being involved) is that the court may hear only one particular type of argument[76] which could thus negatively affect the quality of decision-making.

As far as the contents of the measures ordered by the court is concerned, Rajamani holds that, for example, in the Delhi vehicular pollution case, the court imposed an extremely high cost option without examining whether lower cost alternatives were available.[77] She argues that even though CNG is an environmentally friendly fuel, it is also extremely costly and emits more greenhouse gases than a comparable diesel vehicle. She continues by arguing that notwithstanding the introduction of the CNG programme there still was a 21.3 per cent increase in lung diseases in Delhi and a more than a 20 per cent increase in asthma attacks,[78] thus casting further doubt on the cost-effectiveness of the measures ordered by the court.

Rajamani formulates a similar criticism with respect to the court's ruling in the municipal solid waste management case. She argues that the rules do not contain incentives either to promote recycling

[76] Rajamani (2007), pp. 305–6.
[77] Ibid., p. 308.
[78] Ibid., p. 312.

or to minimize waste since they basically permit the processing of waste through incineration, leading potentially to large emissions of dioxins.[79] This author therefore seriously doubts whether the measures ordered by the court were indeed cost-effective since cheaper alternatives which have not been studied could be available. Rajamani moreover criticizes the fact that the court imposes a relatively high cost environmental solution which may, more particularly in the case of a developing country like India not correspond to the preferences of the citizens but rather to the preferences of the middle-class judiciary of which the Supreme Court in India consists.[80]

9.7 LESSONS FROM THE INDIAN CASE

The particular experience of the Indian case combined with the framework in the literature we addressed above shows that some indicators can be provided on the circumstances under which public interest environmental litigation may be effective; at the same time some indicators can also be given on how to increase the effectiveness of this public interest litigation.

First, given the still-presumed superior informational advantage of regulatory and administrative authorities to set environmental standards in a cost-effective way, an environmental standard setting through the judiciary should only take place as a second-best option when (inter alia as a result of capacity or corruption problems) a standard setting through the regulation fails or is not enforced.

Second, the case of the Supreme Court of India shows that when standard-setting through the judiciary takes place as an alternative for regulation by the executive, guarantees should be provided that the judiciary has the necessary information to set cost-effective standards. One way to do this is to make use of committees that can advise the court. One should, however, be careful that the judiciary does not intervene through its decisions in the functioning of the market, for example, by fixing prices for specific commodities like CNG.

Third, in order to obtain both the guarantee of cost-effective standard setting and the guarantee of effective compliance and enforcement, a high stakeholder involvement in the decision-making

[79] Ibid., pp. 309–10.
[80] Ibid., p. 303.

by the court is indicated, either by having the stakeholders involved in the committees or allowing them to provide information (e.g., as amicus curiae) on various alternative environmental options to reach the goals desired by the court at different costs.

Fourth, reliance on the court and judicial activism via public interest litigation only makes sense in cases where it is clear that problems that occur at the level of the legislator or executive do not occur in the same way with the judiciary. Hence, when public interest litigation is used as a method for environmental standard-setting given the capacity and corruption problems with the executive, judicial independence should guarantee that better results can be achieved by standard-setting through the judiciary.

Fifth, given the fact that a court only reacts in one particular case and in a reactive manner, standard setting as a result of public interest litigation can only be a temporary solution to intervene when the political and legislative system (temporarily) fails. Ideally, the result of judicial activism should be to move the regulatory and administrative authorities to fulfil their task of setting cost-effective environmental standards in the public interest and enforcing them in an adequate way as well. Indeed, in the long run only via standard setting through regulatory authorities can sustainable solutions be achieved.

9.8 JUDICIAL DECISIONS vs EXECUTIVE DECISIONS

Having analysed the cost-effectiveness of some of the environmental cases in which measures were ordered by the Supreme Court of India as a result of public interest litigation, one comes certain to nuanced conclusions: it seems that thanks to these rulings the pollution levels in India indeed substantially decreased, a decrease which probably could not have been realized through the existing legislative or regulatory system. In that sense some of these rulings surely had their environmental effectiveness. On the other hand, the contents of the rulings may not always have been 'optimal' in the economic sense since the alternative imposed by the court may not always have been the lowest cost option to reach the goal set.

A similar type of nuanced conclusion is also presented by Rajamani who studied the public interest environmental litigation in India. She argues that 'in a system in which policy-makers and law-

enforcers are perceived as apathetic, if not corrupt, and politicians are perceived as opportunistic demagogues rather than as visionary leaders,' the Supreme Court of India has performed extremely well as far as providing environmental protection is concerned. This judicial activism is therefore qualified as a 'chemotherapy for the carcinogenic body politic'.[81]

This, therefore, leads to the not surprising conclusion that given the problems at the regulatory level, the Supreme Court of India clearly has filled a gap where regulation failed. However, other criticisms than the lack of cost-effectiveness are made about the public interest environmental litigation, more particularly related to the fact that the judiciary tries to substitute judicial governance for executive governance, when the phenomenon of public interest litigation is not suited to reach sustainable environmental solutions. The judicial decisions are reactive rather than proactive and can even have the negative consequence of restricting the growth of a responsible and independent bureaucracy.[82] Given the inability of the judiciary to provide a long-term judicial oversight and its inability to act proactively, the tendency of the judiciary to position itself as the main protector of the environment[83] can, in the long run be counterproductive since no sustainable solution is achieved.

What can now be learned from this particular case of India with respect to the question of the circumstances under which public interest environmental litigation could play a positive role in environmental policy, particularly in developing countries? Literature on this subject has indicated that if, at a certain level of government, regulatory efforts seem ineffective because of capacity or corruption problems, the option to choose a particular instrument or level of government that is less vulnerable to those capacity or corruption problems is indicated.[84] The result of this literature is that when capacity and corruption problems (which unfortunately occur often in developing countries) are taken into account, a second best solution may be chosen which deviates from the traditional advice provided

[81] Rajamani (2007), pp. 293, 295.
[82] Rajamani (2007), p. 319.
[83] See also Rosencranz and Jackson (2003), p. 249.
[84] See more particularly Ogus (2004, 2006) and Schäfer (2006).

by law and economics scholarship. Apparently the (relative) success story of the public interest litigation in India fits into this strand of literature. Whereas a first best solution for environmental problems would probably be the use of regulation (as inter alia indicated in Shavell)[85]—given the presumed information advantage of regulatory authorities—this first best option may not be achievable in cases where serious capacity or corruption problems exist. In those situations it may be better to look for second-best solutions if other levels of government or instruments are less vulnerable to the mentioned problems.

The case of India shows that where the legislative and administrative bodies are to a large extent unable to implement an effective environmental policy, the judiciary (and more particularly the Supreme Court), where these problems apparently are not encountered, has stepped in to provide this second-best solution. To a large extent this therefore complies with the suggestions made in the mentioned literature.

However, it is equally clear that such a second-best solution can probably only work for a short term in order to trigger regulatory solutions. Indeed, as was indicated above, given the reactive rather than proactive nature of court decisions and given the (still) informational advantage of regulators the interventions by the court may be able to effect, a remedy for apathy of the executive (probably due to corruption problems). However, in the long run the executive should step in to provide more sustainable solutions after the first steps have been taken by the judiciary. Moreover, the Indian case also shows that the judiciary may be lacking the necessary expertise and information to exercise a cost benefit analysis to enable the setting-up of cost-effective environmental standards.

REFERENCES

Mishra, A. and T.C.A. Anant (2006), 'Activism, Separation of Powers and Development', *Journal of Development Economics*, vol. 81, pp. 457–77.

Anant, T.C.A. and J. Singh (2002), 'Economic Analysis of Judicial Activism', *Economic and Political Weekly*, 26 October, pp. 4433–9.

Diwan, S. and A. Rozencranz (2001), 'Environmental Law and Policy in India', New Delhi: Oxford University Press.

[85] Shavell (1984).

Faure, M. and P. Fenn (1999), 'Retroactive Liability and the Insurability of Long-tail Risks', *International Review of Law and Economics*, pp. 487–500.

Faure, M. and G. Skogh (2003), *The Economic Analysis of Environmental Policy and Law: An Introduction*, Cheltenham: Edward Elgar Publication.

Gunningham, N. and P. Grabosky (1998), *Smart Regulation. Designing Environmental Policy*, Oxford: Clarendon Press.

Hayo, B. and S. Voigt (2007), 'Explaining *De Facto* Judicial Independence', *International Review of Law and Economics*, vol. 27, pp. 269–90.

Laffont, J. and M. Meleu (2001), 'Separation of Powers and Development', *Journal of Development Economics*, vol. 64, pp. 129–45.

Kaplow, L. (1986), 'An Economic Analysis of Legal Transition', *Harvard Law Review*, vol. 99, pp. 509–617.

Krishnan, J.K. (2008), 'Scholarly Discourse, Public Perceptions, and the Cementing of Norms: The Case of the Indian Supreme Court and a Plea for Research', *Journal of Appellate Process & Practice*, vol. 9, available at http://ssrn.com/abstract=1003811 (last accessed on 17 December 2009).

Ogus, A. (2004), 'Corruption and Regulatory Structures', *Law and Policy*, vol. 26, pp. 329–46.

——— (2007), 'Regulatory Arrangements and Incentives for Opportunistic Behaviour', in T. Eger, M. Faure, and N. Zhang (eds), *Economic Analysis of Law in China*, Cheltenham: Edward Elgar, pp. 151–63.

Raja, A.V. and F. Xavier (2006), 'Economic Efficiency and Public Interest Litigations (PIL): Lessons from India', paper presented at the second annual conference of the Asian Law and Economics Association, National Law School of India University.

Rajamani, L. (2007), 'Public Interest Environmental Litigation in India: Exploring Issues of Access, Participation, Equity, Effectiveness and Sustainability', *Journal of Environmental Law*, vol. 19 (3), pp. 293–321.

Sathe, S.P. (2001), 'Judicial Activism: The Indian Experience', *Washington University Journal of Law and Policy*, vol. 6, Edward Elgar Publishers, 2003, p. 29.

——— (2008), 'Courts, Media and Civil Society in Regulating the Regulator: Lessons from Delhi Air Pollution Case', working paper, Department of Economics, University of Hyderabad.

Schäfer, H.B. (2000), 'The Bundling of Similar Interest in Litigation, the Incentives for Class Action and Legal Actions Taken by Associations', *European Journal of Law and Economics*, p. 183.

——— (2006), 'Rules versus Standards in Rich and Poor Countries: Precise Legal Norms as Substitutes for Human Capital in Low-income Countries', *Supreme Court Economic Review*, vol. 14, pp. 113–35.

Shavell, S. (1984), 'Liability for Harm versus Regulation of Safety', *Journal of Legal Studies*, vol. 13, pp. 357–74.

Xavier, F. and A.V. Raja (2008), 'Regulatory Failure and Judicial Intervention: Does Public Interest Litigation Help?' *German Working Papers in Law and Economics*, vol. 2008, Berkeley Electronic Press.

10

DISPOSAL RATES, PENDENCY, AND FILING IN INDIAN COURTS

An Empirical Study of the Two States of Andhra Pradesh and Kerala*

R.V. Ramana Murthy and Siddik Rabiyath

10.1 THE GENERAL VIEW AND THE EVIDENCE

Law has long been recognized not only as a reflection of the prevailing forces in a given society, but also as a potential instrument of economic change and progressive development. The literature on law and economic theory exemplifies the role of law and judiciary in the development process.[1] The law and judiciary are central to dispute

* An earlier version of this paper is presented at the 'Senior Scholars' Workshop in Law and Economics', Dept of Economics, University of Hyderabad, India, 28–9 February 2008.

[1] See Thorsten Beck, and Ross Levine (2003), 'Legal Institution and Financial Development', Working Paper No. 3136, World Bank); E. Buscaglia (1999), 'Law and Economics of Development', in B. Bouckeart and De Geest (eds), *Encyclopedia of Law and Economics*, Edward Elgar, Cheltenham; E. Buscaglia and W. Ratliff (2000), *Law and Economics in Developing Countries*, Hoover Institution Press, Stanford-California; M. Chemin (2006), 'Does the Quality of the judiciary Shape Economic Activity? Evidence from a Judicial Reform in India', Department of Economics, *UQAM*; C. Claugue (1997), *Institution and Economic Development: Growth and Governance in Less Developed and Post Socialist Countries*, Jon Hopkins University Press, Baltimore; K.W. Dam (2006), 'The Judiciary and Economic Development', *John M Olin Law and Economics*, Working Paper No. 287, Chicago; K.E. Davis and M.J. Trebilcock (2001), 'Legal Reforms and Development', *Third World Quarterly*, 22(1), pp. 22–6; Hector Fix-Fierro (2003), *Court, Justice and Efficiency: A Socio-Legal Study of Economic Rationality in Adjudication*, Hart publishing, Oregon; T.

resolution, and have a fair influence on efficient resource allocation and exchange.[2] The market economy alone cannot achieve exchange efficiency in the presence of transaction costs and externalities without efficient legal institutions, *a la* Coase.[3] Conversely, inefficient legal institutions, and ambiguity and complexity in law can increase

---

Ginsburg (2000), 'Does Law Matter for Economic Development? Evidence from East Asia', *Law and Society Review*, 1 January 2000; C.W. Gray (1991), 'Legal Process and Economic Development: A Case Study of Indonesia', *World Development*, 19(7), pp. 763–77; F.V. Loon and E. Langerwerf (1990), 'Socioeconomic Development and the Evolution of Litigation Rates of Civil Courts in Belgium, 1835–1980', *Law and Society Review*, 24(3), pp. 283–98; R.E. Messick (1999), 'Judicial Reform and Economic Development: A Survey of the Issues', *World Bank Research Observer*, 14(1), pp. 117–29; D. Mookherjee (2004), 'Legal Institutions and Economic Performance', in *The Crisis in Government Accountability: Essays on Governance Reforms and India's Economic Performance*, Oxford University Press, New Delhi; M. Nakabayashi and T. Okazaki (2007), 'Role of Courts in Economic Development: A Case of Prewar Japan', *CIRJE Discussion Paper* No. F-517, www.e.u-tokyo.ac.jp/cirje/research/dp/2007/2007cf517.pdf (last accessed on 17 December 2009); A. Perry (2002), 'The Relationship between Legal Systems and Economic Development: Integrating Economic and Cultural Approaches', *Journal of Law and Society*, 29(2), pp. 283–307; R.A. Posner (1998), 'Creating a Legal Framework for Economic Development', *World Bank Research Observer*; I.F.I. Shihata (1995a), 'Legal Framework for Development: the World Bank's Role in Legal and Judicial Reform', in R. Malcolm, W.H. Malik, and M. Dakolias, (eds), *Judicial Reform in Latin America and the Caribbean*, World Bank, Washington; (1995b), 'Judicial Reform in Developing Countries and the Role of the World Bank', in R. Malcolm, W.H. Malik and M. Dakolias (eds), *Judicial Reform in Latin America and the Caribbean*, World Bank, Washington; World Bank (2001), 'The Judicial System', in the *World Development Report* 2002, Washington.

[2] See E. Buscaglia (2001), 'The Economic Factors behind Legal Integration: A Jurimetric Analysis of the Latin American Experience', *German Working Papers in Law and Economics*, Paper 1; R.D. Cooter and D. Rubinfeld (1989) 'The Economic Analysis of Legal Disputes and Their Resolution', *Journal of Economic Literature*, 25(3), pp. 231–51; J.W. Hughes and E.A. Snyder (1995), 'Litigation and Settlement under the English and American Rules: Theory and Evidence', *Journal of Law and Economics*; Z. Khan (1999), 'Order without Law: Social Capital, Civil Litigation and Economic Development', *Australian Economic History Review*, 39(3), pp. 172–90; (2000), 'Commerce and Cooperation: Litigation and Settlement of Civil Disputes on the Australian Frontier 1860–1900', *Journal of Economic History*, 60(4), pp. 1088–1119; W.M. Landes (1971), 'An Economic Analysis of Court', *Journal of Law and Economics*, vol. 14, April; M. Rowat, W.H. Malik and M. Dakolias (eds) (1995), *Judicial Reform in Latin america and the Caribbean*, World Bank Technical Paper-280, Washington; W.J. Samuels (1989), 'The Legal Economic Nexus', *George Washington Law Review*, vol. 4, no. 6.

[3] Cooter and Ulen (1995).

the transaction costs. Inordinate legal delays can be symptomatic of such inefficiency. While it is not easy to predetermine the optimal duration for disposal of legal suits, nevertheless, inordinate delays in the disposal of cases can render the justice meaningless.[4] Therefore, disposal of cases in a reasonable time period is arguably essential on pragmatic as well as conceptual grounds.[5] If one considers increased disposal rate as an important criteria for disposal efficiency, increase in the disposal rate in courts is a desired objective for judicial administration.[6] To attain a reasonable degree of disposal efficiency, courts should be staffed by sufficient number of judges, provided with adequate infrastructural facilities, and efficient administrative structures. If an increase in disposal efficiency, inter alia, can decrease pendency, a decrease can cause its obverse. Further, an increase in

[4] See R. Dhawan (1986), *Litigation Explosion in India*, M.N. Tripathi Pvt. Ltd. Bombay; Jiang, Q.Y. (2005), *Court Delay and Law Enforcement in China: Civil Process and Economic Perspective*, Doctoral thesis submitted to the University at Hamburg, Germany; Law commission of India (1978a), *The 77th Report on Delays and Arrears in Trial Courts*, Ministry of Law, Government of India, New Delhi; Law commission of India (1978b), *The 78th Report on the Congestion of under Trial Prisoners in Jails*, Ministry of Law, Government of India, New Delhi; Law commission of India (1979), *The 79th Report on Delays and Arrears in High Courts and other Appellate Courts*, Ministry of Law, Government of India, New Delhi; Law commission of India (1988a), *The 124th Report on High Court Arrears: A Fresh Look*, Ministry of Law, Government of India, New Delhi; Law commission of India (1988b), *The 125th Report on Supreme Court: A Fresh Look*, Ministry of Law, Government of India, New Delhi; Q.M. Maarijuddin (1996), 'Delay in Administration of Justice—An Assessment', *Central India Law Quarterly*, 12(3), pp. 36–56; S.E. Oxner (2003), 'The Quality of Judges', *World Bank Legal Review-Law and Justice for Development*, vol. 1, Kluwer Law International, World Bank; A.A. Zuckerman (1999), *Civil Justice in Crisis*, Oxford University Press, London.

[5] See Government of India (1972), High Court Arrear Committee (Justice Shah Committee Report), Supreme Court of India, New Delhi; Government of India (1990), Report of the Arrears Committee (Malimath Committee Report), Supreme Court of India, New Delhi.

[6] A variant can be seen in E. Buscaglia and M. Dakolias (1999), *Comparative International Study of Court Performance Indicators: A Descriptive and Analytical Account*, World Bank Washington; E. Buscaglia and T. Ulen (1997), 'A Quantitative Assessment of the Efficiency of the Judicial Sector in Latin America', *International Review of Law and Economics*, 17(2); M. Dakolias (1999), *Court Performance around the World*, World Bank Technical Paper No. 430, World Bank; S. Djankov, R. La Porta, F.L. de Silanes, and A. Shleifer (2002), 'Courts: The Lexmundi Project', *Working paper NBER No. 8890*; World Bank (2001).

disposal efficiency can increase efficient out-of-court settlements, under the shadow of law. Correspondingly, a fall in disposal efficiency can discourage the filing of suits in the courts and may lead to inefficient out-of-court settlements. Pendency in courts can thus increase when the rate of filing exceeds the rate of disposal, since filing itself is a function of several factors such as population, national income, rights awareness, etc.

Empirical studies by Buscaglia,[7] Buscaglia and Dakolias[8] and Buscaglia and Ulen[9] pointed out the ill-effects of judicial backlogs, judicial inefficiency, and judicial corruption resulting in the erosion of the confidence of citizens in contractual arrangements that adversely affect economic development. The World Bank recently conducted studies on the significance of judicial and legal reforms among the member countries and argued that there is a definite relation between a country's economic performance and its legal framework, and stressed the need for judicial reform.[10]

Besides this viewpoint of seeing it as a result of institutional and administrative inefficiency, there are models developed in the literature that analyse pendency as a function of utility maximization behaviour by defendants and plaintiffs under alternative legal rules.[11]

The state of research on pendency and related aspects in India is still at a nascent stage and the focus of received studies is still

[7] E. Buscaglia (1996), 'Legal and Economic Integration in Latin America: The Lessons of MERCOSUR', Paper presented at the comparative law and economics meeting in Siena, Italy, and May (1999, 2001).

[8] E. Buscaglia, and M. Dakolias (1996), *Judicial Reforms in Latin American Courts: the Experience in Argentina and Ecuador*, World Bank Technical Paper No. 350, Washington, and Buscaglia and Dakolias (1999).

[9] Buscaglia and Ulen (1997).

[10] World Bank (2001); World Bank (2004), *Initiative in Legal and Judicial Reform*, Washington, Washington DC.

[11] Landes (1971); Cooter and Rubinfeld (1989); R.C. Marshall, M.J. Meurer, and J.F. Richard (1994), 'Litigation Settlement and Collusion', *Quarterly Journal of Economics*, 109(1), pp. 211–39; Hughes and Snyder (1995); L. Vereeck, and M. Muhl (2000), 'An Economic Theory of Court Delay', *European Journal of Law and Economics*, 5(8); A. Torre (2003), 'The Impact of Court Delays on the Prosecutor and the Defendant: An Economic Analysis', *European Journal of Law and Economics*, 2(3); R. Kirstein, and N. Rickman (2004), 'Third Party Contingency Contracts in Settlement and Litigation', *Journal of Institutional and Theoretical Economics*, 160(4), p. 555.

on administrative and institutional aspects.[12] There is a dearth of empirical research on exploring the functional relationships between administrative aspects and legal efficiency. The present paper is an effort in this direction to explore the empirical relationships between disposal, pendency, and filing, based on secondary data, at the macro level on civil litigation for two states, namely, Andhra Pradesh and Kerala. We tried to test five hypotheses drawn on theoretical and empirical studies.

The first hypothesis suggests that increasing pendency has an inverse relation with filing of fresh suits in the long run.[13] This is based on the idea that increasing pendency would increase disposal time, and this would in turn discourage litigants to file fresh suits in court.[14] The second hypothesis is that increased disposal to the proportion of the caseload per judge per annum (disposal rate) *ceteris paribus* would reduce pendency. The third hypothesis is that an increase in both population and income would increase the filing of suits in courts in general. This idea is mainly derived from the empirical works of Toharia, Narayan, and Smith.[15] Toharia argues that 'the caseload of the courts had not kept pace with economic growth.' This impressed us to test the proposition that economic growth has a positive impact on the caseload.[16]

The fourth hypothesis is that increased disposal efficiency would increase fresh filing in courts. This is based on the idea that increased

[12] See R. Dhawan (1978), 'The Supreme Court under Strain: the Challenges of Arrears', M.N. Tripathi Pvt. Ltd. Bombay, 1986. A.K. Hazra (2003), 'The Law and Economics of Disputes Resolution in India', Bookwell, New Delhi. See also various reports of *Law Commission of India.*

[13] Buscaglia (1996).

[14] See Edgardo Buscaglia (1999), Buscaglia, Edgardo and Thomas Ulen (1997).

[15] See in J.J. Toharia (1974) *Cambio Socialy Vida Juridica en Espana*, Edicusa, Madrid; P.K. Narayan and R. Smyth (2006), 'Temporal Causality and the Dynamics of Judicial Appellate Caseload, Real Income and Socio-Economic Complexity in Australia', *Applied Economic*, 23(3), pp. 167–23; R.A. Posner (1999), *The Federal Courts: Challenges and Reform*, Harvard University Press, Cambridge-Massachusetts.

[16] Our case load measurement model is a variant of H.O. Lawson and B.J. Gletne (1980), *Workload Measures in the Court*, National Center for State Courts, Williamsburg; P. Casey (1995), 'Trial Court Performance Standard', National Center for State Courts, Washington DC; available online www.ncsconline.org (last accessed on 4 December 2009).

judicial efficiency would encourage plaintiffs to file suits. And fifth, an increase in the number of judges and judicial expenditure will have a positive effect on the disposal efficiency.

10.2 DEFINITIONS AND DATA SOURCES

The data on judicial statistics is published by the respective offices of Registrars of High Courts, which are reproduced in some states by the Statistical Abstract, *Bureau of Economic Statistics*. The data for the present study is collected from various issues of *Statistical Abstracts*, Govt. of Andhra Pradesh, *Administrative Reports* (Civil and Criminal wing) published by the Government of Kerala, and *Annual Report* of Ministry of Law, Justice and Company Affairs, Government of India. Other variables like national income (NDSP) and population are obtained from *National Accounts Statistics*, Central Statistical Organization, Government of India. Since legal statistics are published by different offices, sometimes there is no uniformity in the data, making inter-state comparison somewhat little difficult. The information on civil suits is available under four heads, namely, 'filing', 'for disposal' (caseload), 'disposed' and 'pendency'. Filing implies filing of fresh suits, usually defined as Newly Instituted (Filed) Litigations.[17] For disposal, or in other words the 'caseload', is the cumulative sum of cases filed in the current year plus cases pending from the previous years. Cases 'disposed' include cases on which verdicts are declared, cases are transferred without trial, dismissed on admission, compromised, referred for arbitration, etc. Pendency of cases denotes cumulative pendency at the beginning of the year plus fresh suits filed minus cases disposed. We can also say that pendency constitutes the current year's total undisposed litigation at the end of the year. The litigation statistics presented here constitutes the litigations under the purview of High Courts and subordinate courts in the states of Kerala and Andhra Pradesh during 1980–2000, the choice of the states being purely incidental.[18] The choice of period of study is based on availability

[17] Newly instituted cases are the fresh litigations that are either suits, appeals, or petitions. Cases for disposal refers to the action of disposing, that is, settling or definitely dealing with (a case); being disposed of; transferring into new hands or getting done with. See, Constitution of India, Sch. 7, List III entry 41.

[18] Civil suits are filed at *munsif* (local) courts as well as District Courts and appeals are filed at High Courts and above.

of comparable data. As said earlier, we considered only civil suits filed in lower and high courts in Kerala and Andhra Pradesh.

10.3 TRENDS IN FILING, CASELOAD, DISPOSAL, AND PENDENCY IN KERALA AND ANDHRA PRADESH

The growth of the filing of civil cases in the two states depicts two contrasting trends. In the case of Andhra Pradesh (AP), the filing of civil cases is continuously increasing over time, as can be expected with any process of development, while the filing in Kerala too is increasing when considered without the state-specific civil miscellaneous cases[19] (once we include civil miscellaneous litigation,[20] filing shows a declining trend in comparative as well absolute number of cases[21]). There is a great decline in the number of miscellaneous cases in Kerala since 1986–2000, the reason being that several state acts were passed periodically that were specific to the contexts that trigger certain unusual litigation.[22] The growth rates of the four variables namely, filing, caseload, disposal, and pendency in cases over the same period in Andhra Pradesh and Kerala, are given in the Table 10.1.

[19] Civil miscellaneous cases constitute a major portion of civil litigation in Kerala. But this disaggregation is not found in Andhra Pradesh's data set published by government agency. Therefore, Kerala has two separate columns as total for the analysis—one for the civil cases excluding the civil miscellaneous cases and the second includes civil miscellaneous cases in the total civil litigation.

[20] Civil miscellaneous litigation constitutes miscellaneous cases under state-specific special enactments like Succession Certificate Act, Guardian and Ward Act, to sue informa pauper, Probation and Administration Act, Building Lease and Rent Control Act, Lunacy Act, Marriages Act, Vellala Act, Nayar Act, Ezhava Act, Village Panchayath Courts Act, Chitty Act, Agricultural Debt Relief Act, Agrarian Relation Act, other Original Petitions, Land Acquision References, Motor Vehicle Act, Electricity O Ps, Miscellaneous Petitions connected with E Ps, and Miscellaneous petitions other than EPs.

[21] We were unable to get satisfactory explanation for such a steep fall of civil cases in Kerala from field inquiries with the Registrar's office, Kerala High Court. There is one explanation—that miscellaneous cases belong to the State-specific Acts, as mentioned in the above foot note, which are promulgated under specific contexts, the cases therefore show behaviour related to those specific contexts.

[22] Many times the rules and circumstances pertaining to the applicability of acts come with some delay, and meanwhile lots of suits get filed and judiciary may delay verdicts until the publication of the rules. This could cause piling up of filing, delayed/suspended disposal and increased pendency.

TABLE 10.1: Growth Rates of Litigation in Kerala and Andhra Pradesh (1980–1 to 2000–1)

States	*Andhra Pradesh*	*Kerala*	
Variables	*Total*	*Total without civil miscellaneous cases*	*Total with civil miscellaneous cases*
Newly instituted (Fresh Filing)	1.48	2.95	−2.96
For disposal (caseload)	2.78	4.32	1.32
Disposed cases	1.91	3.22	−3.34
Pendency	3.42	5.10	5.41

Source: *Statistical Abstract*, GOAP and *Administration Report* (civil wing) GOK.

The trend growth rate of caseloads has increased significantly in both the states over the last twenty years. In the case of Andhra Pradesh and Kerala, the trend growth rates[23] are 2.78 and 4.32 per cent per annum respectively. The inclusion of miscellaneous cases dampens the growth of caseload to 1.32 per cent in Kerala during the period, perhaps due to the steep decline in the filing in the category. In spite of this, the growth rate of caseloads seems to be higher in Kerala, both in absolute and relative number, with or without including civil miscellaneous cases.[24]

The growth rates of disposed cases in Kerala is 3.22 per cent per annum over the period of 1980–1 to 2000–1, while it is 1.91 per cent in Andhra Pradesh, which is still less than the caseload growth of 4.32 and 2.78 per cent respectively, in Kerala and Andhra Pradesh. The disposal scenario seems to be grim in case of Kerala when miscellaneous cases are included, where it falls to a negative growth rate of −3.34 per cent. Though the disposal rate is higher than the filing rate in both Kerala and Andhra Pradesh, it is not sufficient to reduce the caseload growth, given the large absolute number of cases filed, resulting in an accumulation of pending cases. In Kerala's case, there is a relatively steep fall of disposal of miscellaneous cases and because of this, the caseload in the state has gone up more. The pendency rates in the two states are 5.41 and 3.42 per

[23] Trend Growth.

[24] The filing, caseload, disposal and pendency statistics would be presented in the appendix Table 10.1.

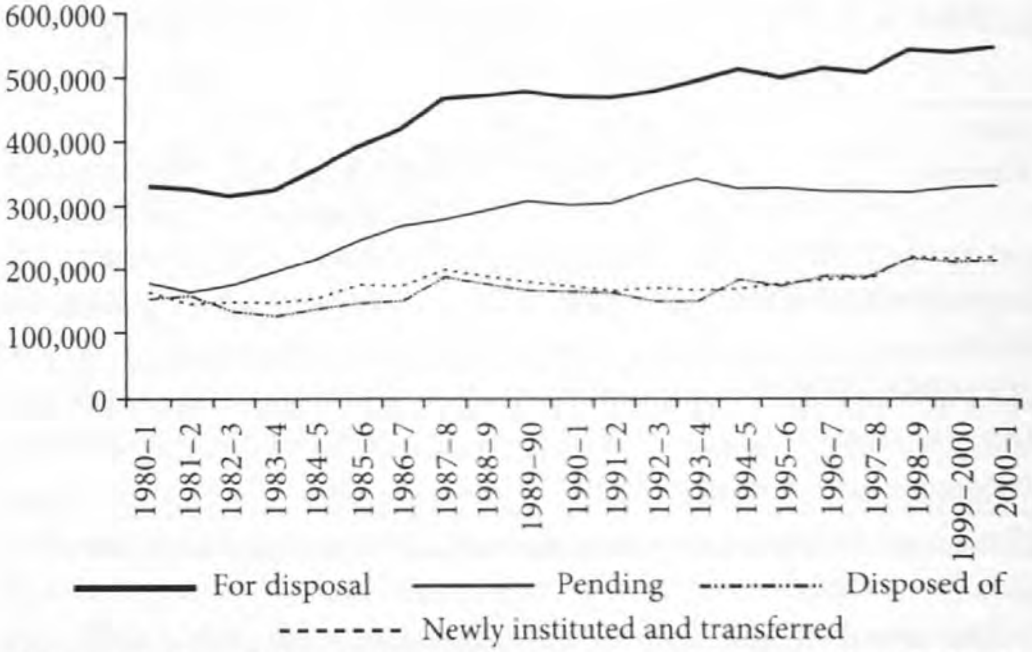


FIGURE 10.1: Litigation Trends in Andhra Pradesh (1980–1 to 2000–1)

Source: Author.
Note: Total Litigation (Original+Small Cause Suits/Petition+Regular and Miscellaneous Appeals+Execution and Insolvency Petitions) in all Courts of Andhra Pradesh.

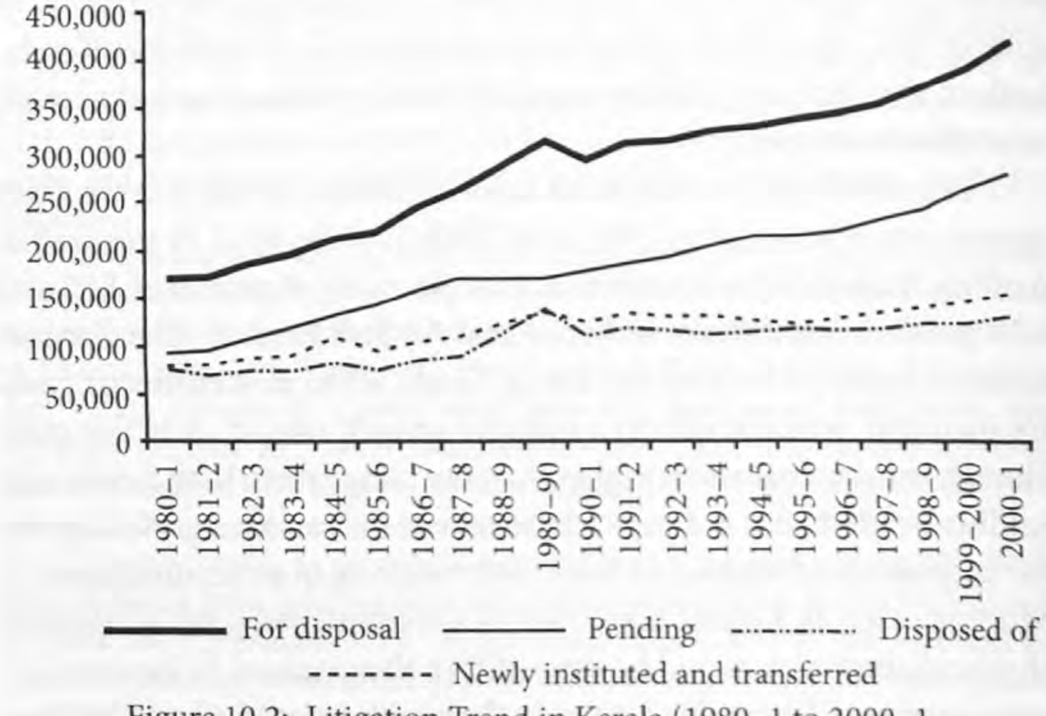


Figure 10.2: Litigation Trend in Kerala (1980–1 to 2000–1, without miscellaneous cases)

Source: Author.
Note: Total Litigation (Original+Small Cause Suits/Petitions+Regular and Miscellaneous Appeals+Execution and Insolvency Petitions) in all Courts of Kerala.

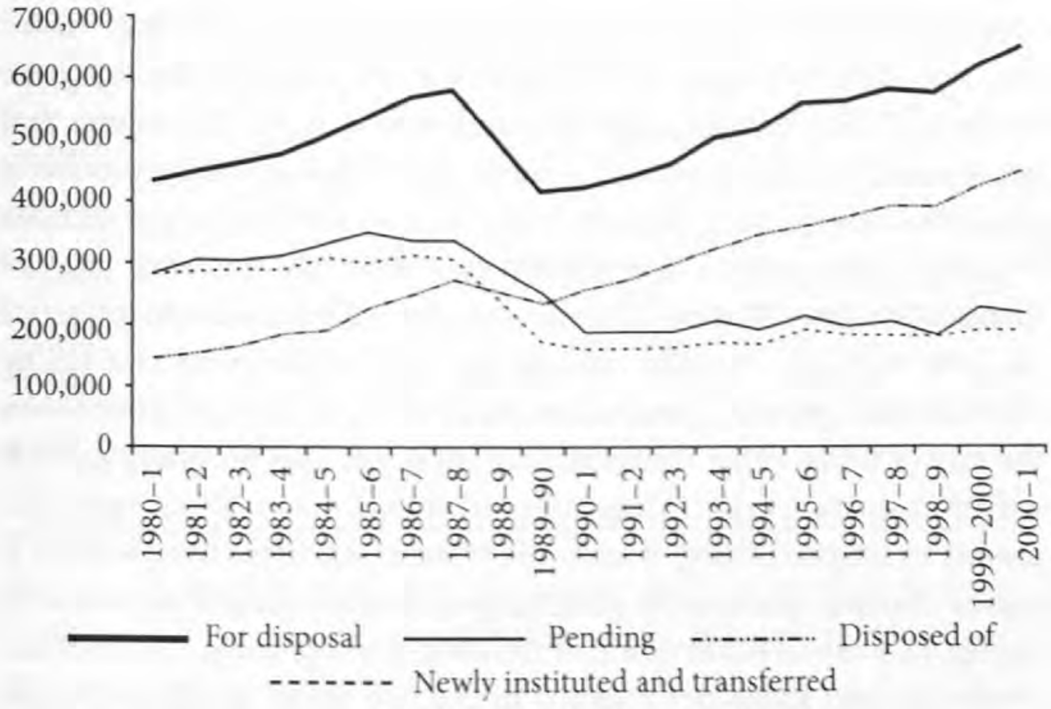


FIGURE 10.3: Litigation Trends in Kerala (1980–1 to 2000–1, with miscellaneous cases)

Source: Author.
Note: Total Litigation (Original+Small Cause Suits/Petitions+Regular and Miscellaneous Appeals+Execution and Insolvency Petitions+Other Miscellaneous Petitions) in all Courts of Kerala.

cent per annum in Kerala and Andhra Pradesh respectively. This larger pendency in Kerala is only due to the factor of miscellaneous cases and without that the pendency rates are more or less same in the two states. However, there is an improvement in the disposal rates in the nineties in Andhra Pradesh, though not in Kerala,[25] which seems to be a serious matter. Figures 10.1 to 10.3 give litigation status in Andhra Pradesh and Kerala. We will further probe this matter by looking at the proportion of disposals and filings to the caseload, which perhaps give some clues to understand these trends further.

10.3.1 Filing and Disposal—Some Further Long-run Factors of Determination

The growth of civil litigation can depend on a number of factors. Generally, factors such as population growth, net state domestic product,

[25] See Table 10.1 in the appendix.

number of judges, judicial expenditure,[26] rates of urbanization,[27] inter alia, can influence rates of civil litigation. We examine the influence of these factors on filing and disposal of civil suits. We assume that population, net state domestic income, and urbanization can positively affect the rate of filing of fresh cases, and an increase in the number of judges and judicial expenditure can positively influence rates of disposal. But then it is not clear how the rate of disposals can influence the rate of filing. Possibly, one can say that in the event of a falling disposal rate, growing pessimism about the judiciary can drive down the rate of filing, other things remaining same. This proposition holds with the common belief, 'filing follows disposal', that is, disputing parties prefers to litigate (filing) if and only if their case is resolved sooner. It means that in a system with a high disposal of cases, the filing would be higher. In a casual observation of trends in average filing, disposal, and pendency (per judge per annum) in the two states of AP and Kerala [see Figures 10.4 to 10.6] we can see the above proposition.

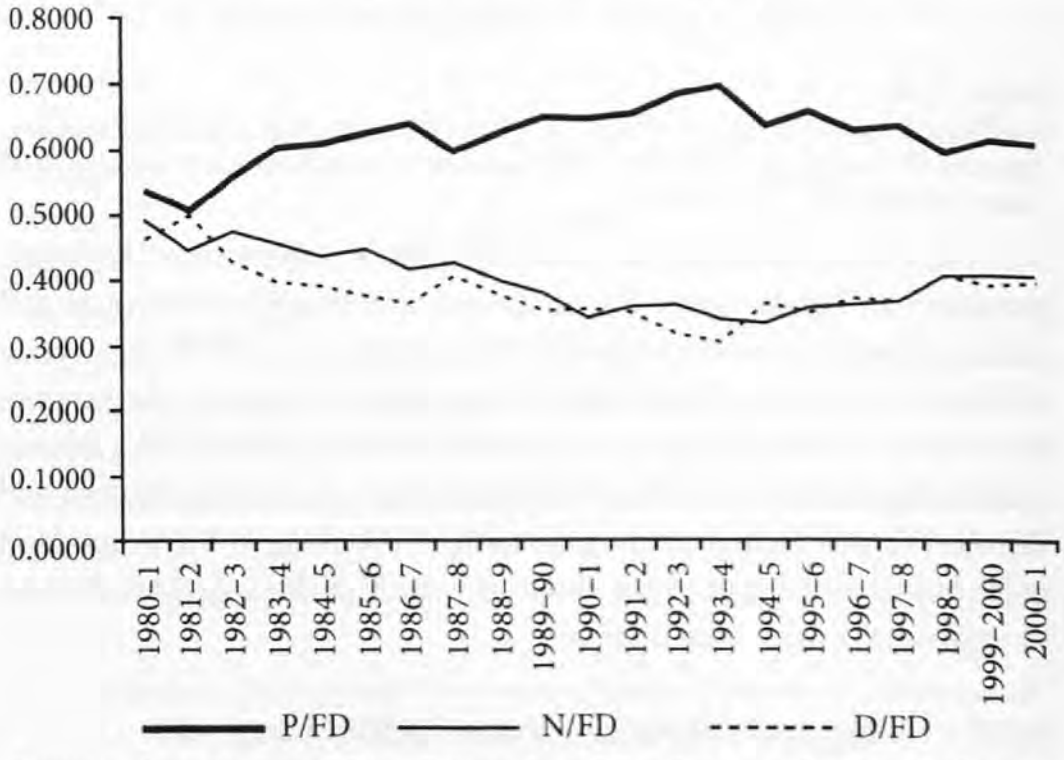


FIGURE 10.4: Average Filing, Disposal, and Pendency to the Caseload in AP

Source: Author.

[26] See appendix Table: 10.2 for the absolute figure.

[27] In the study by Narayan and Smyth (2006) urbanization took as a proxy for the determinants of litigation.

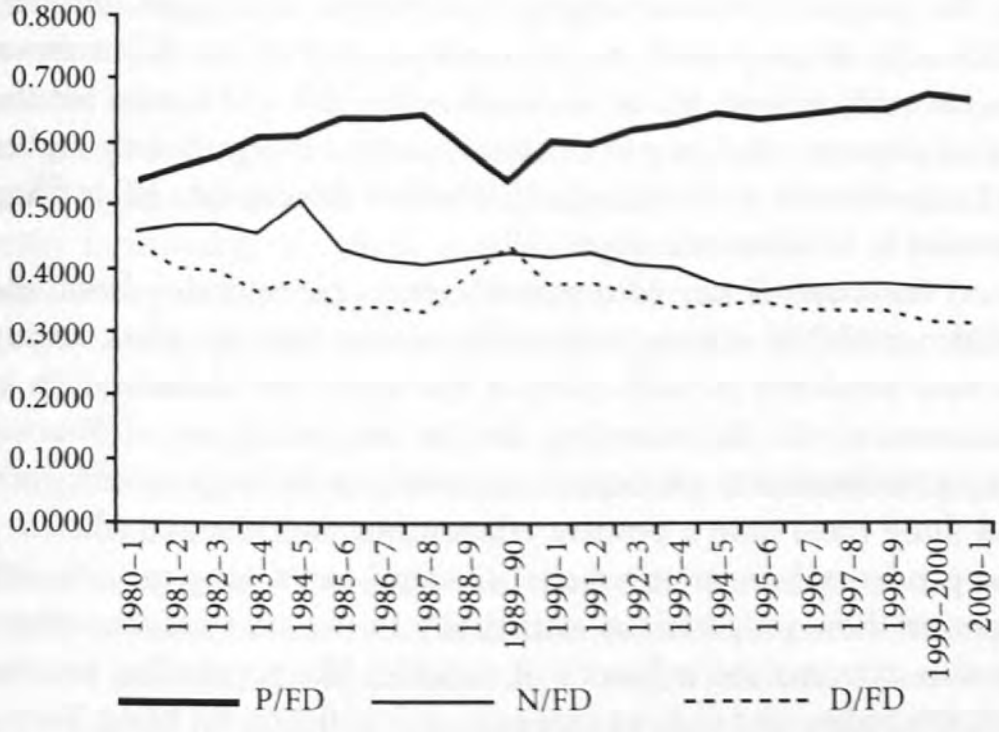


FIGURE 10.5: Average Filing, Disposal, and Pendency to the Caseload in Kerala (Excluding Other Civil Miscellaneous Cases)

Source: Author.

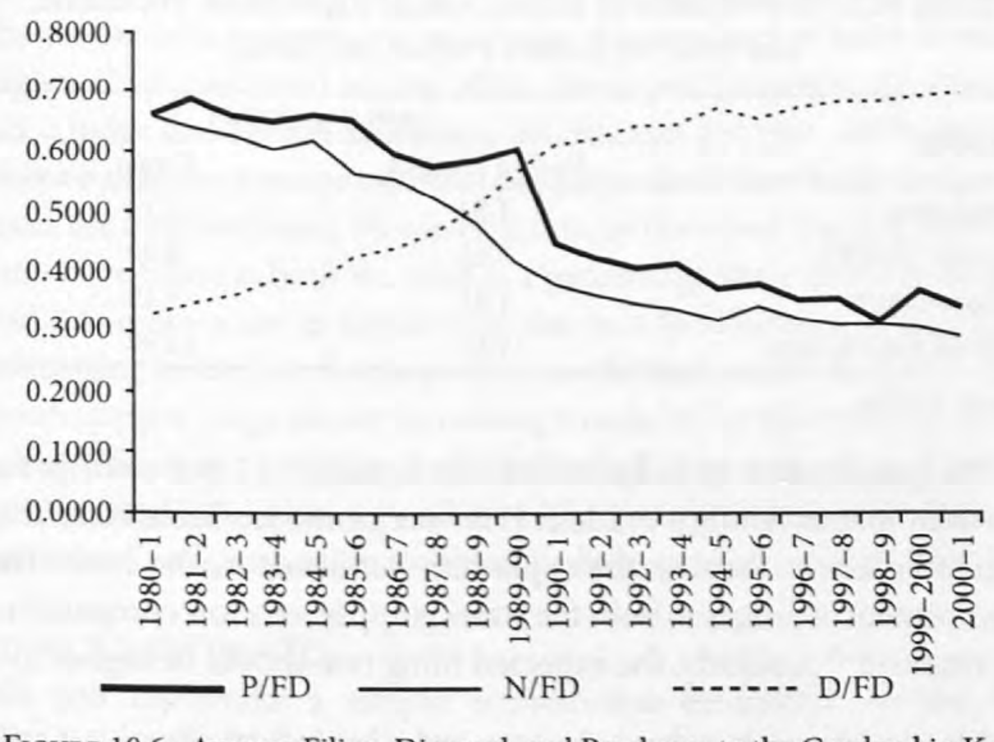


FIGURE 10.6: Average Filing, Disposal, and Pendency to the Caseload in Kerala (Including Other Civil Miscellaneous Cases)

Source: Author.

The graphs, therefore, suggest that the rate of disposal (disposal efficiency) in proportion to the caseload (D/FD) has fallen down considerably in both the states, Andhra Pradesh and Kerala, but the fall of disposal efficiency to the total caseload is significantly higher in Kerala than in Andhra Pradesh. Whether this caused a fall in filing remains to be examined empirically.

As the disposal rate (disposal efficiency) curve comes down, the pendency (P/FD) curves predictably goes up over the years, largely because pendency includes filing of new suits. This observation is in consonance with the reasoning that the decreasing rate of disposal leads to an increase in pendency; that *ceteris paribus* 'disposal efficiency and filing cases have a positive relationship' and 'disposal efficiency and pendency have an inverse relationship' in the long run. We will examine these propositions statistically in the next section, where we also examine the influence of variables like population, income (NSDP), judges and judicial expenditure (public) on the filing. Before moving to the estimation, we will examine the trends in growth rates of these variables.

TABLE 10.2: Growth Rates of Judges, Judicial Expenditure, Population, and NSDP in Andhra Pradesh and Kerala

Variables	*1980–1 to 2000–1*	
	Andhra Pradesh	*Kerala*
Population	1.84	1.12
Income (NSDP)	5.62	4.90
Judges/Courts	1.81	1.12
Judicial Expenditure	14.10	14.90

Source: Author.

The growth rate of judges/courts in Kerala (1.12 per cent) is far less than that in Andhra Pradesh (1.81 per cent), see Table 10.2. This is far too less to dispose the exploding pendency in the state. The appointment of judges in both the states is apparently low compared to the caseload.[28] Similarly, the expected filing rate should be higher in a

[28] In our interviews, judges, advocates, and other judicial officers in Kerala also expressed the same view. They were of the view that even when the pendency is at its peak, the authorities never bother to fill the posts of those judges who are on leave, backlog vacancies.

state with a higher rate of growth of population, other things remaining constant. If so we can expect a higher rate of filing in Andhra Pradesh than in Kerala. When it comes to the judicial expenditure in Kerala and Andhra Pradesh it shows a positive uptrend over the period of analysis. Though growth in judicial expenditure is desirable for the better functioning of judicial system, the positive effects of judicial expenditure also depend on the efficacy of its use. When there is no commensurate growth in the number of judges, this alone may not be a sufficient factor.

10.3.2 Average Filing, Caseload, Disposal, and Pendency per Judge in the two States

When we examined the number of cases filed per judge in the two states, we found similar trends in conformity with the earlier findings. Table 10.3 shows it clearly. In the case of Andhra Pradesh, the caseload has increased from 654.71 cases in 1980–81 to 768.50 cases in 2000–01, whereas in Kerala the caseload has increased from 492.77 to 848.94 in the same period, but if civil miscellaneous litigation is included, it goes up from 1251.25 cases to 1322.96 in the same period. Besides the comparison between the two states, this caseload in India is much higher than compared to any other developed countries (in case of US, a judge disposes on an average 40–50 cases per year, see Posner).[29] These figures on average caseload per judge show how badly judges in India are overburdened. However, it is to be observed that the disposal rates have fallen in both the state as a percentage share to the caseload, and it is more acute in Kerala over the past two decades. It would be interesting to explore the issue of causes behind such a steep fall. The pendency per judge shows increasing trends in the two states. A faster average filing per judge in Andhra Pradesh over a stagnant disposal rate resulted in increased pendency. The same trend is seen when we consider Kerala's situation (without including miscellaneous cases).

10.4 OLS ESTIMATES

We will undertake a simple econometric estimation to test the hypotheses that we have proposed in Section 10.1. We will also test the influence of factors such as population, net state domestic

[29] Posner (1999).

Table 10.3: Litigation Per Judge in all Courts

Year	*Andhra Pradesh**				*Kerala**				*Kerala***			
	Filed	*For disposal*	*Disposed of*	*Pending*	*Filed*	*For disposal*	*Disposed of*	*Pending*	*Filed*	*For disposal*	*Disposed of*	*Pending*
1980–1	321.80	654.71	303.13	351.58	226.92	492.77	215.08	263.99	824.26	1251.25	824.30	413.24
1981–2	282.56	634.96	314.68	320.28	224.90	481.41	193.04	269.96	865.46	1267.00	807.80	440.80
1982–3	290.78	612.37	262.00	342.77	235.59	502.70	197.72	290.07	831.15	1263.48	791.61	456.96
1983–4	288.73	631.24	249.91	381.32	234.23	514.02	188.55	313.08	822.58	1263.95	763.06	488.50
1984–5	275.90	631.83	246.91	384.92	276.94	543.57	209.46	332.82	855.91	1292.99	801.33	490.38
1985–6	303.58	678.56	255.00	423.57	239.66	557.04	185.53	357.16	887.80	1361.13	766.54	580.24
1986–7	289.15	696.58	252.21	444.37	247.73	598.75	204.06	382.74	826.20	1391.54	769.42	610.17
1987–8	327.78	770.69	310.87	459.82	264.03	645.37	215.68	418.33	825.66	1435.56	752.06	672.13
1988–9	312.93	779.50	294.33	485.17	291.42	695.91	270.28	410.89	704.56	1205.33	577.54	613.05
1989–90	296.12	783.14	275.97	507.17	315.51	741.24	321.55	402.25	591.52	978.14	406.92	553.77
1990–1	250.73	735.64	261.21	474.44	285.30	677.33	254.10	409.83	430.70	970.41	365.67	591.34
1991–2	266.15	750.48	260.69	489.78	292.72	689.63	258.58	412.87	417.35	990.09	356.82	615.09
1992–3	267.80	743.86	236.12	507.74	277.56	677.89	245.70	420.83	406.48	1002.88	345.77	645.75
1993–4	260.93	768.67	234.04	534.63	281.65	702.07	237.07	443.43	446.91	1092.26	366.63	704.06

(Contd.)

(Table 10.3 contd.)

Year	*Andhra Pradesh**				*Kerala**				*Kerala***			
	Filed	*For disposal*	*Disposed of*	*Pending*	*Filed*	*For disposal*	*Disposed of*	*Pending*	*Filed*	*For disposal*	*Disposed of*	*Pending*
1994–5	262.28	792.79	286.71	506.08	266.72	704.45	240.49	456.45	409.07	1104.09	353.57	743.00
1995–6	250.83	708.38	246.94	466.16	256.14	704.90	244.41	450.47	443.43	1173.91	399.39	764.50
1996–7	260.45	725.29	269.04	456.25	266.82	714.44	240.06	460.64	414.87	1174.55	374.19	786.63
1997–8	262.89	720.43	263.06	457.37	278.04	739.55	244.29	485.25	425.58	1213.76	376.74	827.00
1998–9	317.14	785.06	318.30	466.77	294.83	783.16	258.38	513.45	382.38	1214.62	375.58	827.70
1999–2000	305.41	754.46	294.30	461.00	296.36	795.73	251.77	538.23	463.33	1267.94	385.60	876.60
2000–1	307.27	768.50	301.75	466.75	312.70	848.94	264.84	570.50	449.72	1322.96	387.72	921.63

* It includes Original +Small Cause suits/petitions + Regular and Miscellaneous appeals + Execution and Insolvency Petitions.

** It includes Original +Small Cause suits/petitions + Regular and Miscellaneous appeals + Execution and Insolvency Petitions + other civil miscellaneous cases

Source: Statistical Abstract of Andhra Pradesh, Government of Andhra Pradesh from 1980–81 to 2000–01 for Andhra Pradesh data and Administration Report (Civil Wing), Government of Kerala from 1980–81 to 2000–01.

product, number of judges, and judicial expenditure on the filing of cases in the two states for the period of 1980–1 to 2000–1. We will use a simple two variable ordinary least squares as well as multi variable OLS estimation.[30]

We mainly test five hypotheses, based on the preceding discussion. These are, first, increasing pendency would have an inverse relation with filing of fresh suits in the long run. This is based on the idea that the piling up of litigation increases the courts' congestion. As a result, the expected duration of dispensation would be higher, which would discourage the parties to litigate. Similarly, net expected costs also determine the decision of plaintiffs to sue. Once the pending cases go up, the probability of rational selection to go for litigation goes down and vice versa. Secondly, increased disposals would reduce pendency. Thirdly, the increase in population and income can increase in the filing of suits in courts. Also increased disposals would increase fresh filing in courts. And finally, increase in judges and judicial expenditure would have a positive effect on the disposal of cases.

First, we would consider our simple OLS regression pendency on filing, and then disposal on pendency. Then we move on to the multiple variable OLS estimation of number of judges and judicial expenditure on disposal (see Table 10.4).

The coefficient of pendency with respect to filing is found to be negative in the case of Kerala during the last two decades, even when we included the miscellaneous cases. This suggests that for one per cent increase in pendency, filing will decline by 0.82 (without including civil miscellaneous cases) and when they are included, it declines by 0.516 in Kerala. Therefore, in Kerala's case, we accept the hypothesis that increasing pendency would discourage litigation, which does not seem to be the case in A.P, where the coefficient shows a significantly positive sign with an elasticity of 0.59.

Our second hypothesis is that the disposal efficiency will have an inverse relation with pendency.[31] Since, in Andhra Pradesh we have

[30] Owing to a severe problem of multicolinearity, we confined the exercise to simple OLS estimation in some cases and used multiple regression estimation otherwise.

[31] $\Rightarrow P_i = f\,(D_i) + \mu_i^{24} \qquad (i = 1, 2, \ldots, n) \qquad (2)$

TABLE 10.4: Two Variables OLS Regression Results

Section A

Period	*1980–1 to 2000–1*		
Controlled variable	*Control variable*		
Filing	*Pending*		
States	*Andhra Pradesh*	*Kerala*	
	Total	*Total without miscellaneous cases*	*Total with miscellaneous cases*
Constant	4.635 (1.344)	23.54 (7.515)	18.63 (12.838)
Coefficient	0.592** (2.176)	−0.829* (−3.771)	−0.516* (−4.292)
Std. error	0.272	0.219	0.120
r-squared	0.743	0.923	0.492
Ad *r*-squared	0.713	0.923	0.465

Section B

Controlled variable	*Control variable*		
Pendency	*Disposal*		
	Total	*Total without miscellaneous cases*	*Total with miscellaneous cases*
Constant	0.407 (0.119)	−358.2 (−0.007)	22.89 (9.944)
Coefficient	1.006* (3.553)	−0.423* (−4.484)	−0.882* (−4.2711)
Std. error	0.283	0.094	0.187
r-squared	0.399	0.984	0.538
Ad *r*-squared	0.367	0.982	0.514

* indicates 1 per cent and ** indicates 5per cent level of significance
Figures in parentheses shows 't'-statistic
Estimation done by using log values

not noticed any drastic fall in the disposal efficiency of the courts, we do not contemplate this result.

The estimated result shows the inverse relationship between pendency and disposal in the case of Kerala and not in AP. The elasticity in Kerala is more with the inclusion of miscellaneous cases (−0.882) than without miscellaneous litigation (−0.423) in each unit change in the disposal efficiency. In the case of AP, despite a decline in disposal efficiency, the filing has still gone up, perhaps due to other factors.

Now we turn to our third hypothesis that population, income, and disposal efficiency influence the filing of suits in Kerala and Andhra

TABLE 10.5: Multiple OLS Regression Results ($\Rightarrow F_t = f(Pop_t, NSDP_t, D_t) + \mu_t$)

Period	*1980–1 to 2000–1*														
States	*Andhra Pradesh*					*Kerala*									
Controlled	*Total*					*Total without miscellaneous cases*					*Total with miscellaneous cases*				
variable	*Control variables*			*Constant & lag*		*Control variables*			*Constant & lag*		*Control variables*			*Constant & lag*	
Filing	*Pop*	*NSDP*	*D*	*C*	*Ar. (1)*	*Pop*	*NSDP*	*D*	*C*	*Ar. (1)*	*Pop*	*NSDP*	*D*	*C*	*Ar. (1)*
Coefficient	0.83	–0.06	0.50*	–7.45	0.48	1.41*	–0.03	0.50*	–17.6	0.38	0.64	–0.06**	0.86*	–7.49	0.003
t-statistic	1.63	–1.03	2.22	–0.93	1.74	2.31	–1.18	4.68	–1.94	1.69	0.96	–1.93	8.42	–0.64	0.01
Std. error	0.51	0.06	0.22	7.97	0.27	0.61	0.10	0.10	9.08	0.22	0.66	0.03	0.10	11.62	0.25
P value	0.12	0.31	0.04	0.36	0.10	0.03	0.25	0.00	0.07	0.10	0.34	0.07	0.00	0.52	0.98
R-squared			0.803					0.965					0.941		
Ad R2			0.750					0.956					0.925		
D-W stat			1.634					1.965					1.990		
F-stat			15.30					106.07					60.160		

* indicates 1 per cent, ** 5 per cent and *** 10 per cent levels of significance.

Pradesh.[32] From the estimation it can be seen that (Table 10.5), though the expected sign of coefficient of filing with respect to population is positive, it is not found to be statistically significant. The coefficient of filing over NSDP has perverse negative sign. However, the coefficient of filing over disposal efficiency bore positive sign in case of both the states and statistically significant. From these regressions, we infer that the decision to litigate is more dependent on the disposal efficiency (which determines the pendency and via filing of suits) than on the other two factors. Now we will examine the issue whether judges' number improves the disposal efficiency.

10.5 CAN THE NUMBER OF JUDGES INCREASE DISPOSAL EFFICIENCY?

It is crucial to note that disposal rate depends on the number of judges and courts. Normally, it is expected that an increase in the number of judges will increase the disposal of cases by courts over time. But whether it can improve the average disposal per judge is not very clear. We have seen earlier in Tables 10.2 and 10.3 that the number of judges in the two states have gone up in both absolute number and growth rate. However, if the increase in judges is less in relation to the level of filing and pendency, then it may not make any dent on pendency. However, it should certainly have a positive impact on the disposal of cases in all courts. The study therefore examines the common proposition that an increase in the number of judges has a positive impact on disposals. The regression results are given in Table 10.6.

In the case of Andhra Pradesh, the results show (see Table 10.6) that the coefficient of the number of judges with respect to disposal is positive and significant, the elasticity being 1.088. Correspondingly in Kerala, the elasticity is higher at 2.002 (excluding the miscellaneous cases), but it becomes negative when miscellaneous category cases are included (elasticity being –1.928). Since the miscellaneous cases in Kerala seem to be related to some specific context, the negative coefficient seems to be because of those special reasons. The positive sign without miscellaneous supports the general case. To conclude, the growth of judges does matter for the disposal efficiency, which

[32] $\Rightarrow F_i = f(Pop_i, NSDP_i, D_i) + \mu_i^{25} \quad (i = 1, \ldots, n)$ (3)

TABLE 10.6: Two Variables OLS Regression Results
($\Rightarrow D_1 = f(J_1) + \mu_1$)

Controlled variable	*Control variable*		
Disposal	*Judges*		
States	*Andhra Pradesh*	*Kerala*	
	Total	*Total without miscellaneous cases*	*Total with miscellaneous cases*
Coefficient	1.088* (3.570)	2.002* (4.309)	−1.928* (−5.905)
Constant	5.032 (2.560)	−0.636 (−0.224)	23.964 (12.113)
Std. error	0.304	0.464	0.326
r-squared	0.668	0.840	0.647
Ad r-squared	0.629	0.821	0.628

* indicates 1 per cent level of significance; Figure in parentheses shows 't'-statistic

of course improves the confidence of litigants and they might relay on the institutional set-up to maximize their utility.[33] Therefore, we conclude that the increase in the number of judges is very important to increase the disposal rate, at the same time the disposal efficiency of judges perhaps also can be increased.

10.5 IMPACT OF JUDICIAL EFFICIENCY

To conclude, the present study is a modest effort in search of empirical evidence on some of the issues centring around judicial efficiency and its impact in a law and economics framework. We tried testing some hypotheses relating to behaviour of filing of civil suits, caseload, disposal, and pendency for the two states of Kerala and Andhra Pradesh. The lack of standardized data poses some difficulty in making smooth comparisons between states. Nevertheless, the study succeeded in validating some important notions related to efficiency of courts and their impact. The results of the study could confirm the hypotheses that a falling disposal rate of cases in courts would increase pendency rates; and increasing pendency rates reduce the rate of filing. This means that increasing pendency in courts and the consequent delays could reduce the confidence of litigants

[33] However, an increase in the number of judges perhaps has not been enough to counter the falling disposal efficiency of the judges; the negative sign is, therefore, plausible.

in filing cases in courts. We also found that increased disposal rate increases the filing rate, other things remaining constant. The availability of judges has a decisive impact on disposal efficiency and pendency. Our study supports the general notion that an increase in the number of judges, given the caseload in India, is extremely important to increase the efficiency of legal institutions. We did not find statistically significant factors such as population and state domestic product to have any influence on filing. The trend in filing of civil cases in absolute rate in Andhra Pradesh showed a continuous rise over the period. The pendency rates in Kerala are found to be relatively higher. Yet this growth in terms of the proportion of filing cases to the total caseload is showing a downward trend. This lends credence to the view that decreasing disposal efficiency could bring down the rate of filing of suits. The major reason seems to be a lack of sufficient number of judges for the kind of caseload that exists in the Indian courts, rather than the disposal rate as such.

REFERENCES

Beck, T. and R. Levine (2003), 'Legal Institution and Financial Development', Working Paper No.3136, World Bank.

Buscaglia, E. (1996), 'Legal and Economic Integration in Latin America: The Lessons of MERCOSUR', Paper presented at the comparative law and economics meeting in Siena, Italy, May.

—— (1999), 'Law and Economics of Development', in B Bouckeart and De Geest (eds), *Encyclopedia of Law and Economics*, Edward Elgar, Cheltenham.

Buscaglia, E. (2001), 'The Economic Factors behind Legal Integration: A Jurimetric Analysis of the Latin American Experience', *German Working Papers in Law and Economics*, Paper 1.

Buscaglia, E., and M. Dakolias (1996), *Judicial Reforms in Latin American Courts: the Experience in Argentina and Ecuador*, World Bank Technical Paper No. 350, Washington.

—— (1999), *Comparative International Study of Court Performance Indicators: A Descriptive and Analytical Account*, World Bank Washington.

Buscaglia, E., and T. Ulen (1997), 'A Quantitative Assessment of the Efficiency of the Judicial Sector in Latin America', *International Review of Law and Economics*, 17(2).

Buscaglia, E. and W. Ratliff (2000), *Law and Economics in Developing Countries*, Hoover Institution Press, Stanford-California.

Casey, P. (1995), 'Trial Court Performance Standard', National Center for State Courts, Washington DC; available online www.ncsconline.org (last accessed on 4 December 2009).

Chemin, M. (2006), 'Does the Quality of the judiciary Shape Economic Activity? Evidence from a Judicial Reform in India', Department of Economics, *UQAM*.

Claugue, C. (1997), *Institution and Economic Development: Growth and Governance in Less Developed and Post Socialist Countries*, Jon Hopkins University Press, Baltimore.

Cooter, R.D., and D. Rubinfeld (1989) 'The Economic Analysis of Legal Disputes and Their Resolution', *Journal of Economic Literature*, 25(3), pp. 231–51.

Cooter, Robert and Thomas Allen (1995), *Law and Economics*, Addison–Wesley, New York.

Dakolias, M. (1999), *Court Performance around the World*, World Bank Technical Paper No. 430, World Bank.

Dam, K.W. (2006), 'The Judiciary and Economic Development', *John M Olin Law and Economics*, Working Paper No. 287, Chicago.

Davis, K.E., and M.J. Trebilcock (2001), 'Legal Reforms and Development', *Third World Quarterly*, 22(1), pp. 22–26.

Dhawan, R. (1978), *The Supreme Court under Strain: The Challenges of Arrears*, M.N. Tripathi Pvt Ltd, Bombay.

—— (1986), *Litigation Explosion in India*, M.N. Tripathi Pvt Ltd, Bombay.

Djankov, S., R. La Porta, F.L. de Silanes, and A. Shleifer (2002), 'Courts: The Lexmundi Project', *Working paper NBER No. 8890.*

Hector Fix-Fierro (2003), *Court, Justice and Efficiency: A Socio-Legal Study of Economic Rationality in Adjudication*, Hart publishing, Oregon.

Friedman, L.M. (1990), 'Opening the Time Capsule: A Progress Report on Studies of Courts over Time', *Law and Society Review*, 21(2).

Ginsburg, T. (2000), 'Does Law Matter for Economic Development? Evidence from East Asia', *Law and Society Review*, 1 January 2000.

Government of Andhra Pradesh, *Statistical Abstract*, Annual Issues from 1980–1 to 2001–2, Hyderabad.

Government of India (1972), *High Court Arrear Committee (Justice Shah Committee Report)*, Supreme Court of India, New Delhi.

—— (1990), *Report of the Arrears Committee (Malimath Committee Report)*, Supreme Court of India, New Delhi.

——, *Annual Report of the Ministry of Law, Justice and Company Affairs*, various issues, Ministry of Law, New Delhi.

Government of Kerala, *Administration Report on Judiciary (Civil Wing)*, Annual Issues form 1980-1 to 2001-2, Thiruvananthapuram.

——, *Administration Report on Judiciary (Criminal Wing)*, Annual Issues from 1980-1 to 2001-2, Thiruvananthapuram.

Gray, C.W. (1991), 'Legal Process and Economic Development: A Case Study of Indonesia', *World Development*, 19(7), pp. 763-77.

Gujarati, D.N. (2005), *Basic Econometrics*, Tata McGraw-Hill, New Delhi.

Hazra, A.K. (2003), *The Law and Economics of Disputes Resolution in India*, Bookwell, New Delhi.

Hughes, J.W., and E.A. Snyder (1995), 'Litigation and Settlement under the English and American Rules: Theory and Evidence', *Journal of Law and Economics*.

Jiang, Q.Y. (2005), *Court Delay and Law Enforcement in China: Civil Process and Economic Perspective*, Doctoral thesis submitted to the Universitat Hamburg, Germany.

Khan, Z. (1999), 'Order without Law: Social Capital, Civil Litigation and Economic Development', *Australian Economic History Review*, 39(3), pp. 172-90.

—— (2000), 'Commerce and Cooperation: Litigation and Settlement of Civil Disputes on the Australian Frontier 1860-1900', *Journal of Economic History*, 60(4), pp. 1088-1119.

Kirstein, R., N. and Rickman (2004), 'Third Party Contingency Contracts in Settlement and Litigation', *Journal of Institutional and Theoretical Economics*, 160(4), p. 555.

Laeven, L., and G. Majnoni (2005), 'Does Judicial Efficiency Lower the Cost of Credit', World Bank Series No. 3159.

Landes, W.M. (1971), 'An Economic Analysis of Court', *Journal of Law and Economics*, vol. 14, April.

Lawson, H.O., and B.J. Gletne (1980) *Workload Measures in the Court*, National Center for State Courts, Williamsburg.

Law Commission of India (1978a), *The 77th Report on Delays and Arrears in Trial Courts*, Ministry of Law, Government of India, New Delhi.

—— (1978b), *The 78th Report on the Congestion of under Trial Prisoners in Jails*, Ministry of Law, Government of India, New Delhi.

—— (1979), *The 79th Report on Delays and Arrears in High Courts and Other Appellate Courts*, Ministry of Law, Government of India, New Delhi.

—— (1988a), *The 124th Report on High Court Arrears: A Fresh Look*, Ministry of Law, Government of India, New Delhi.

—— (1988b), *The 125th Report on Supreme Court: A Fresh Look*, Ministry of Law, Government of India, New Delhi.

Loon, F.V., and E. Langerwerf (1990), 'Socioeconomic Development and the Evolution of Litigation Rates of Civil Courts in Belgium, 1835–1980', *Law and Society Review*, 24(3), pp. 283–98.

Maarijuddin, Q.M. (1996), 'Delay in Administration of Justice—An Assessment', *Central India Law Quarterly*, 12(3), pp. 36–56.

Marshall, R.C., M.J. Meurer, and J.F. Richard (1994), 'Litigation Settlement and Collusion', *Quarterly Journal of Economics*, 109(1), pp. 211–39.

Messick, R.E. (1999), 'Judicial Reform and Economic Development: A Survey of the Issues', *World Bank Research Observer*, 14(1), pp. 117–29.

Mookherjee, D. (2004), 'Legal Institutions and Economic Performance', in *The Crisis in Government Accountability: Essays on Governance Reforms and India's Economic Performance*, Oxford University Press, New Delhi.

Nachane, D.M. (2006), *Econometrics: Theoretical Foundations and Empirical Perspectives*, Oxford University Press, New Delhi.

Nakabayashi, M., and T. Okazaki (2007), 'Role of Courts in Economic Development: A Case of Prewar Japan', *CIRJE Discussion Paper* No. F-517, www.e.u-tokyo.ac.jp/cirje/research/dp/2007/2007cf517.pdf (last accessed on 17 December 2009).

Narayan, P.K., and R. Smyth (2006), 'Temporal Causality and the Dynamics of Judicial Appellate Caseload, Real Income and Socio-Economic Complexity in Australia', *Applied Economic*, 23(3), pp. 167–23.

Oxner, S.E. (2003), 'The Quality of Judges', in *World Bank Legal Review—Law and Justice for Development*, vol. I, Kluwer Law International, World Bank.

Perry, A. (2002), 'The Relationship between Legal Systems and Economic Development: Integrating Economic and Cultural Approaches', *Journal of Law and Society*, 29(2), pp. 283–307.

Pistor, K., Wellon, P.A., *et al.* (1999), *The Role of Law and Legal institutions in Asian Economic Development*, Oxford University Press, Hong Kong.

Posner, R.A. (1998), 'Creating a Legal Framework for Economic Development', *World Bank Research Observer.*

—— (1999), *The Federal Courts: Challenges and Reform*, Harvard University Press, Cambridge-Massachusetts.

Rabiyath, S. (2007), *Litigation, Disposal Efficiency and Pendency: A Comparative Analysis of Kerala and Andhra Pradesh*, unpublished M Phil thesis, submitted to the University of Hyderabad, India.

Rowat, M., W.H. Malik, and M. Dakolias (eds) (1995), *Judicial Reform in Latin America and the Caribbean*, World Bank Technical Paper-280, Washington.

Samuels, W.J. (1989), 'The Legal Economic Nexus', *George Washington Law Review*, 4(6).

Shihata, I.F.I. (1995a), 'Legal Framework for Development: The World Bank's Role in Legal and Judicial Reform', in R. Malcolm, W.H. Malik, and M. Dakolias (eds), *Judicial Reform in Latin America and the Caribbean*, World Bank, Washington.

—— (1995b), 'Judicial Reform in Developing Countries and the Role of the World Bank', in R. Malcolm, W.H. Malik, and M. Dakolias (eds), *Judicial Reform in Latin America and the Caribbean*, World Bank, Washington.

Toharia, J.J. (1974) *Cambio Socialy Vida Juridica en Espana*, Edicusa, Madrid.

Torre, A. (2003), 'The Impact of Court Delays on the Prosecutor and the Defendant: An Economic Analysis', *European Journal of Law and Economics*, 2(3).

Vereeck, L., and M. Muhl (2000), 'An Economic Theory of Court Delay', *European Journal of Law and Economics*, 5(8).

World Bank (2001), 'The Judicial System', in the *World Development Report* 2002, Washington.

—— (2004), *Initiative in Legal and Judicial Reform*, Washington, Washington DC.

Zuckerman, A.A. (1999), *Civil Justice in Crisis*, Oxford University Press, London.

Appendix Table A.1: Civil Litigation: Status of Kerala and Andhra Pradesh from 1980-1 to 2000-1

Year	*Total Number of Newly Instituted Suits, Appeals and Petitions*			*Cumulative Total (caseload)*			*Disposed of*			*Cumulative pendency at the end*		
	*AP**	*Kerala**	*Kerala***	*AP**	*Kerala**	*Kerala***	*AP**	*Kerala**	*Kerala***	*AP**	*Kerala**	*Kerala***
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
1980–1	164118	77835	282721	333900	169020	429179	154595	73771	282736	179305	90547	141741
1981–2	145235	79388	305509	326369	169936	447251	161747	68144	285155	164623	95297	155601
1982–3	150332	85990	303371	316596	183486	461171	135456	72166	288936	177210	105877	166792
1983–4	149271	88303	310113	326352	193786	476511	129201	71083	287675	197143	118033	184166
1984–5	156159	107451	332093	357614	210905	501680	139750	81272	310915	217864	129135	190267
1985–6	176379	93948	348019	394245	218359	533562	148153	72726	300483	246092	140008	227454
1986–7	174644	101074	337089	420736	244289	567750	152334	83256	313923	268402	156156	248950
1987–8	198637	107195	335218	467039	262019	582836	188387	87565	305338	278652	169841	272885
1988–9	189637	120357	290983	472376	287409	497803	178364	111626	238526	294014	169696	253190
1989–90	180635	133144	249621	477714	312802	412773	168341	135692	171719	309373	169749	233693
1990–1	159967	123537	186494	469340	293286	420187	166649	110025	158333	302691	177457	256051
1991–2	166342	130847	186557	469051	308266	442570	162930	115585	159498	306111	184552	274943

(Contd.)

(Table A.1 contd.)

Year	Total Number of Newly Instituted Suits, Appeals and Petitions			Cumulative Total (caseload)			Disposed of			Cumulative pendency at the end		
	AP*	Kerala*	Kerala**	AP*	Kerala*	Kerala**	AP*	Kerala*	Kerala**	AP*	Kerala*	Kerala**
1	2	3	4	5	6	7	8	9	10	11	12	13
1992–3	172194	127956	187386	478305	312506	462327	151826	113266	159398	326479	194003	297692
1993–4	167778	129840	206027	494257	323656	503532	150488	109289	169015	343769	204422	324572
1994–5	169959	124556	191034	513728	328980	515608	185786	112308	165117	327942	213161	346980
1995–6	177838	121666	210628	502239	334827	557608	175083	116095	189709	330506	213972	363139
1996–7	185177	127539	198307	515683	341501	561436	191288	114750	178864	324395	220186	376007
1997–8	186388	132627	203003	510783	352765	578962	186510	116526	179705	324273	231462	394480
1998–9	219777	139751	181248	544050	371216	575731	220579	122470	178027	323471	243374	392332
1999–2000	219588	144622	226104	542459	388316	618756	211602	122864	188175	331457	262654	427783
2000–1	220928	153223	220362	552554	415981	648249	216959	129770	189981	335595	279544	451601

Source: Compiled from the various issues of Statistical Abstract of Andhra Pradesh, Administrative Reports (Civil & Criminal wing) published by the Government of Kerala and Annual Report of Ministry of law, Justice and Company Affairs

Notes: * It includes Original +Small Cause suits/petitions + Regular and Miscellaneous appeals + Execution and Insolvency Petitions

** It includes Original +Small Cause suits/petitions + Regular and Miscellaneous appeals + Execution and Insolvency Petitions + other civil miscellaneous cases

Appendix Table A.2: Status of the Exogenous Variables such as Population, NSDP, Judges, and Judicial Expenditure in Kerala and Andhra Pradesh

Year	*Population*		*NSDP*		*No. of Judges/ Courts*		*Judicial expenditure (Rs)*	
	AP	*Kerala*	*AP*	*Kerala*	*AP*	*Kerala*	*AP*'	*Kerala*
1980–1	53077000	25357000	732395	382273	510	343	98357708	51337728
1981–2	54229000	25699000	850960	377481	514	353	130506861	57892098
1982–3	55414000	26046000	866043	386851	517	365	147772179	66008056
1983–4	56627000	26398000	902230	371168	517	377	184760940	76008721
1984–5	57869000	26754000	874901	394158	566	388	180710739	81413306
1985–6	59145000	27115000	930296	408636	581	392	227473358	99918523
1986–7	60443000	27481000	904974	399297	604	408	274283591	121020731
1987–8	61765000	28114000	1027403	416563	606	406	297216918	120507309
1988–9	63120000	28402000	1202883	458410	606	413	317562400	121020731
1989–90	64503000	28693000	1298577	489236	610	422	463723212	174706690
1990–1	65912000	28987000	1358012	526234	638	433	555780167	215366595
1991–2	67218000	29378000	1434166	536466	625	447	613331265	229423491
1992–3	68433000	29775000	1395455	575161	643	461	590731853	243480387

(Contd.)

(*Table A.2 contd.*)

Year	Population		NSDP		No. of Judges/ Courts		Judicial expenditure (Rs)	
	AP	Kerala	AP	Kerala	AP	Kerala	AP*	Kerala
1993–4	69637000	29879000	1554312	634665	643	461	752450956	307244804
1994–5	70763000	30132000	5456383	2590792	648	467	776292328	368598017
1995–6	71805000	30388000	5795084	2694747	709	475	887002869	412461376
1996–7	72767000	30646000	6195503	2802645	711	478	1026139427	464192259
1997–8	73639000	30906000	6032120	2863315	709	477	1087782145	516346315
1998–9	74409000	31168000	6803640	3060430	693	474	1395934536	627922690
1999–2000	75067000	31432000	7090408	3278501	719	488	1708471661	789972319
2000–1	75604000	31699000	7707692	3396268	719	490	1854198742	817022249

Source: CSO, Government of India; Statistical Abstract, Government of Andhra Pradesh Administrative Reports (civil & criminal wing), Government of Kerala.
Note: * Expected (minimum) judicial expenditure during the period.

(Table 6.2 contd.)

Year	Population		NSDP		No. of Judges/ Courts		Judicial expenditure (Rs)*	
	AP	Kerala	AP	Kerala	AP	Kerala	AP	Kerala
1993–4	68417000	29870000	1152314	654005	348	441	75868998	30759994
1994–5	70163000	30123000	1422223	2330?94	368	329	71853520	38069043
1995–6	71669000	30384000	1516694	[illegible]	366	373	86065869	41757136
1996–7	72741000	30640000	1749444	[illegible]	721	278	[illegible]	[illegible]
1997–8	73619000	30990000	1697120	[illegible]	789	427	100703195	[illegible]
1998–9	74549000	31180000	1960240	[illegible]	489	376	[illegible]	[illegible]
1999–2000	75467000	31412000	2056408	[illegible]	716	446	[illegible]	[illegible]
2000–1	75608000	31649000	2170692	[illegible]	729	490	[illegible]	[illegible]

Source: CSO, Government of India; Statistical Abstract, Government of Andhra Pradesh; Administrative Reports (civil & criminal wing), Government of Kerala.

Note: * Expected (minimum) judicial expenditure during the periods.

VI

THE IMPACT OF NON-LEGAL INSTITUTIONS ON ECONOMIC DEVELOPMENT

11

THE LIMITED ROLE OF LAW IN ECONOMIC DEVELOPMENT

Thomas S. Ulen*

Why some nations have grown and prospered and others have not is one of the most remarkable puzzles of modern scholarship. And considering how long the topic has fascinated scholars, it is surprising how little we truly know about the subject. We know what distinguishes modern growth from episodes of sporadic well-being in the past: modern growth is sustained and, therefore, cumulative, whereas in the past there have been civilizations that prospered for a time—think of Spain in the sixteenth and seventeenth centuries; the ancient Mayan, Aztec, and Incan empires; and those who developed Angkor Wat. What we do know about the sustained growth that characterizes the last two centuries is when and where it began: in the early nineteenth century in northern Europe—most notably in the United Kingdom and the Netherlands.

But there is no consensus on much else about modern economic growth and development. For instance, while there are, as we shall see, many theories of the causes of growth, none of those theories commands universal favour. Nor is there much agreement about what

* I would like to thank Thomas Eger of the University of Hamburg and Raja Angara of the University of Hyderabad for their invitation to participate in this symposium. The discussions with the participants at the Asia-Link Senior Scholar's Conference in Hyderabad in February, 2008, were invaluable in revising this paper. I also want to thank Amber Evans and Brian Doxey of the University of Illinois College of Law Classes of 2008 and 2009 for their marvellous research assistance.

should be done to help the most distressed, the poorest countries in the world to begin the process of growth.[1] And there is great controversy about why some areas of the world—central Asia, sub-Saharan Africa, and much of the Arab world—seem to be nearly immune to modern growth. Nor do we fully understand why some countries that were once relatively prosperous, such as Argentina and Spain, lost much of their wealth and then seem to have begun to regain prosperity.

The search for explanations of what causes growth and how growth can become the standard human condition has ranged far and wide and recently has looked to law as a possibly central causal factor. In this essay I want to discuss the contention that law is (or might be) a central factor in explaining economic growth. My view is that law is important, but not central. Bad law (including no law) can certainly stifle growth, or even reverse it, but the positive case for law as a central force in development is not, I argue, strong.

The essay proceeds in the following steps. In section 11.1 I briefly review the most recent patterns of economic growth and some theories of economic growth, and identify some recent issues in development. Section 11.2 considers the role that law in its many forms might play in fostering development and prosperity. A concluding section 11.3 summarizes the argument.

11.1 PATTERNS OF RECENT ECONOMIC GROWTH

Economic growth has long fascinated scholars. Arguably, the modern science of economics began—in Adam Smith's *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776)—with a treatise on economic growth. Since then, the profession's concern with explaining and fostering growth and development has been episodic. We appear to be in a phase now in which economists (and others) are devoting a great deal of time to the subject. Before we turn to a summary of those efforts, let me first sketch the very long-term patterns of economic growth and then turn to consideration of the very recent record.

[1] See Paul Collier (2007), *The Bottom Billion: Why the Poorest Countries are Failing and What can be done*, Oxford University Press, Oxford, for thoughtful proposals for helping the poorest one-sixth of humanity.

11.1.1 The Long View

Modern economic growth, by which I mean 'sustained economic growth', began in the early nineteenth century (or possibly the mid-eighteenth century) in northern Europe, principally in the United Kingdom and the Netherlands.[2] There had been no discernible increase in human well-being during the first millennium of the modern era, and the best estimate is that average per capita income increased by 50 per cent during the 800 years between AD 1000 and AD 1800. There were differences in the average per capita income across the world in AD 1800, but those differences were not that substantial.[3] The world's richest economy in roughly 1800, the United Kingdom, had an average per capita income that was approximately four times that of the poorest region in the world, Africa.[4]

Between 1820 and 2000, the average per capita income and total world output of goods and services increased substantially. For the world as a whole, average per capita income during this period increased by nine-fold. But some countries prospered far more than that. In the United States, average per capita income increased by twenty-five times; in Western Europe, fifteen times. For the world as a whole (world GDP, if you will), the total value of all goods and services increased by a factor of almost 50.[5] We do not yet fully understand why some countries grew faster than others.

An important implication of this broad, long term view of economic growth is that the phenomenon of increasing societal well-being is modern. For the vast stretch of human history, human beings did not have a reasonable expectation that their standard of living or per capita income would increase over the course of their lifetimes, nor that their children or grandchildren could anticipate easier, healthier, wealthier, or longer lives than they lived.

[2] Angus Maddison dates the beginning of modern growth from 1820. See Angus Maddison (2001), *The World Economy: A Millennial Perspective/Historical Statistics*, OECD, Paris and *The Contours of the Word Economy 1–2030 A.D.: Essays in Macro-Economic History*, Oxford University Press, Oxford (2007).

[3] See Maddison (2007a) and Jeffrey Sachs (2006), *The End of Poverty: Economic Possibilities for Our Time*, Penguin Books, New York.

[4] Sachs (2006).

[5] Maddison (2001), and Sachs (2006).

Increasingly since the early nineteenth century, the notion of growth in well-being and per capita income has come to be a reasonable expectation for humans to have. And that expectation is not just reasonable in the developed economies; it has also become a reasonable expectation of those in the poorer, developing countries. And in some instances the expectation is not just of gradual increases in well-being or increases that will accrue to one's children and grandchildren but of rapid and palpable increases for each generation. Some economies are growing at annual rates in the order of 8–10 per cent. Growth in Gross Domestic Product (GDP) at that pace implies that GDP will double each seven to nine years. Supposing that population growth is not too great, this rate of increase in GDP suggests that average per capita income could double every decade or fifteen years. As health improves and life spans increases, this means that individuals could see their annual incomes double or even quadruple during a single lifetime. That is the sort of demonstrable improvement in matters that creates strong expectations. And those expectations can and do give rise to societal disruptions and political pressures.[6]

To put the matter succinctly, the 180-year-long period of modern economic growth began with all countries roughly in equal conditions of well-being. The richest country in the world had an average per capita income in 1820 that was about 90 per cent of that in a modern African country. The ratio between the average per capita income in the wealthiest country in the world in 1820 and the poorest was on the order of four to one. And average life expectancy in Japan and Western Europe was approximately forty years.

But by 2000, the picture was very different. Approximately one-sixth of humanity had risen to a high-income status—levels of well-being hardly imaginable at the beginning of the nineteenth century. Approximately two-thirds of humanity had risen to middle-class status, with comfortable but not extravagant levels of well-being. And approximately one-sixth of humanity is stuck in extreme poverty,

[6] To my knowledge, no one has written about how the experience of modern economic growth has changed the average person's expectations for his or her life's accomplishments and how those changes then give rise to political action. But the topic is a potentially important one and ought to command scholarly attention.

stuck, that is, at the same levels of well-being that characterized the entire world at the beginning of the modern period.

11.2.2 The Scholarly Literature on the Causes of Growth

When modern economists, beginning with Roy Harrod, Evsey Domar, and Robert Solow, first turned their attention to economic growth, they characterized the growth process as one that was technical and, at root, driven by improvements in the technology of production. Growth was a matter of getting the appropriate ratio of capital to labour in the production process and of using a society's resources efficiently. In the famous and innovative Solow attempt to model growth and to test the applicability of the model's predictions to the record of twentieth century US growth, the vast bulk of growth—well more than 50 per cent—was attributable to the 'residual'—that is, to the growth of 'total factor productivity,' the accumulated improvement in the efficiency with which resources were organized for production (as opposed to growth occurring through increases in the quantities of such factors of production as land, labour, and capital).[7] But this clearly vital engine of growth was not explained within the model. Technological change was exogenous to the model.

The policy implications of the Harrod-Domar and Solow models of economic growth were seized upon in the 1960s and 1970s to guide, aid, and advise to the developing countries. Whether correctly, the central recommendation that advisors gave to the developing world was to increase the amount of productive capital equipment, such as dams, roads, factories, and the like, so as to raise the capital-labour ratio to the levels that would sustain higher growth rates. Because the developing world did not have sufficient domestic savings to invest in capital accumulation, the developed world filled this 'financing gap' through aid that was to be invested in massive capital-accumulation projects.

The result of this (perhaps misguided) application of the Harrod-Domar and Solow models to development was grim failure. There was no discernible growth in the countries that received aid; nor was growth demonstrably faster in those countries that received more aid;

[7] Robert M. Solow (1956), 'A Contribution to the Theory of Economic Growth,' *Q. J. Econ.*, 70(65), and (1957) 'Technical Change and the Aggregate Production Function,' *Rev. Econ. & Stat.*, 3(312).

nor was there any other empirical indication that the aid based on the 'financing gap' model was helpful.[8]

The lack of progress of development in the poor countries through the 1980s gave rise to searches for better theories of growth and better methods of helping the poor nations. Among scholars there was great interest in two strands of thought. The first was the elaboration of 'endogenous growth' (or 'new growth theory') models. The innovation of those models was to bring technical change within the model so that real factors in an economy determined the rate of technical change and, through that rate and other real factors, the nation's growth rate.[9] Although these new growth models were a tremendous scholarly advance, they have not had a deep impact on the policy of advising or aiding developing economies.

The second strand of academic thought on economic growth that gave rise to excitement was the interest in 'institutions' as a fundamental determinant of national growth and development. Douglass North had long been associated with the 'new institutional economics,' which holds (among other things), that the legally and otherwise-determined rules that govern economic activity have a profound influence on the ability of an economy to prosper.[10] For example, the mechanisms by which a society encourages and rewards risk-taking will have a deep effect on investments that might be necessary for growth. If a society discourages economic risk-taking

[8] The story is marvellously told in William Easterly (2001), *The Elusive Quest for Growth: Economists' Adventures and Misadventures in the Tropics*, MIT Press, Cambridge, Mass.

[9] Paul Romer (1986), 'Increasing Returns and Long-Run Growth,' *J. Pol. Econ.*, 94(1002), and Romer (1990), 'Endogenous Technological Change', *J. Pol. Econ.*, 98(S71) . For a nontechnical discussion, see Paul Romer (2007), 'Economic Growth', in David R. Henderson (ed.), *The Concise Encyclopedia of Economics*, Liberty Fund, Indianapolis, USA.

[10] Douglass C. North (2005), *Understanding the Process of Economic Change*, Princeton, NJ, and Douglass C. North and Robert Paul Thomas (1976), *The Rise of the Western World: A New Economic History*. See also Stephen L. Parente and Edward C. Prescott (2000), *Barriers to Riches*. In his wonderful survey of economic thinking about growth, Elhanan Helpman concludes that the technical models had not provided the explanatory power and guidance for practical policymaking that developing countries desire. But, he wrote, the interest in institutions as a bedrock of growth holds great promise for both explaining and directing growth. See Helpman (2004), *The Mystery of Economic Growth*.

or does not have appropriate safety nets to catch those who take risks and fail, then that society may not have sufficient investment in entrepreneurial activity. In theory, economists could identify the institutions that are necessary for a nation to grow and prosper, see if a given nation has those institutions, and if it does not, suggest means by which it might acquire those vital institutions.[11]

One of the most interesting strands to emerge from the institutional literature is that of Daron Acemoglu, Simon Johnson, and James A. Robinson. In a series of insightful articles,[12] they have argued that there is persuasive theoretical and empirical evidence that those nations that have strong legal and social institutions have prospered and that those nations that do not have strong and stable institutions have not prospered. Consider the phenomenon they characterize as the 'reversal of fortune.'[13] The starting point for the article is the observation that in the sixteenth century there were some colonies of European countries that were extremely wealthy (such as some Caribbean islands) and others that were very poor (such as North America). But by the late twentieth century, these roles had been reversed: the former colonies that had been relatively wealthy in the

[11] The 'Washington consensus', which I shall discuss shortly, was the policy implementation of this view. Dani Rodrik of the Kennedy School of Government, Harvard University, has been a consistent champion of the view that institutions are central to economic prosperity. See (April, 2004) 'Getting Institutions Right' (available at http://ksghome.harvard.edu/~drodrik/ifo-institutions%20 article%20_April%202004_.pdf) (last accessed on 21 October 2009) and (2004), 'Institutions Rule: The Primacy of Institutions Over Geography and Integration in Economic Development' (with Arvind Subramanian and Francesco Trebbi) *J. Econ. Growth*, 9(131) (2004). See also Michael Spence (2008), Chair, Commission on Growth and Development, *The Growth Report: Strategies for Sustained Growth and Inclusive Development*, available at http://www.growthcommission.org/index.php?option=com_content&task=view&id=96&Itemid=169. (last accessed on 21 October 2009).

[12] See, for example, Daron Acemoglu, Simon Johnson, and James A. Robinson (2005), 'The Rise of Europe: Atlantic Trade, Institutional Change and Growth', *Am. Econ. Rev.*, 95(546) and (2001), 'The Colonial Origins of Comparative Development: An Empirical Investigation,' *Am. Econ. Rev.*, 91(1369). See also Acemoglu and Robinson (2005), *Economic Origins of Dictatorship and Democracy* and Acemoglu (forthcoming, 2009), *Introduction to Modern Economic Growth*.

[13] Acemoglu, Johnson, and Robinson (2002), 'Reversal of Fortune: Geography and Institutions in the Making of the Modern World Income Distribution,' *Q. J. Econ.*, 117(1231).

sixteenth century were now relatively (and in some cases absolutely) poor and those that had been poor in the sixteenth century are now wealthy. What explains this reversal of fortune?

According to Acemoglu, Johnson, and Robinson, the key difference between the two groups of colonies is the presence of institutions that would facilitate long-run growth. The colonies that were wealthy in the sixteenth century were 'extractive': they had substantial deposits of valuable natural resources that the European colonizers extracted so as to get those resources onto ships and into markets in Europe. Because these colonies were valued only to the extent that they had valuable natural resources and those resources were finite, the colonizers did not have a long-term interest in the colony. As a result, there was no reason for the colonizers to invest in establishing and nurturing stable and valuable institutions. For example, there was nothing to be said for establishing representative and responsive governance structures in the colonies. The inhabitants of the colonies were to be exploited, not encouraged to be self-governing. Nor was there much to be said for encouraging widespread and secure property interests or the other institutions of a well-functioning civil society.

As a result, when the natural resources ran out, these extractive colonies began a long slide into poverty with no stable institutions to arrest that slide or to provide alternative sources of economic growth.

Contrast this account of the history of extractive colonies with that of what Acemoglu, Johnson, and Robinson call 'settler colonies'.[14] In the sixteenth century these colonies had no immediately valuable natural resources that could be extracted for transportation to markets in Europe. In order to become valuable to their European masters, these colonies would require longer-run development. In particular, they would need to become populated with people who could develop the colony so that it would become valuable to the mother country. So, these colonies sought to encourage emigration from the mother country (and other European countries). And the settlers would be more eager to come to the extent that there were attractive prospects of success in the colony. This fact, argue Acemoglu, Johnson, and Robinson, explains why these colonies developed stable, attractive,

[14] Acemoglu, Johnson, and Robinson (2005).

citizen-empowering institutions. And those institutions served the long-run interests of the colonists so well that over the long haul those colonies became wealthy.[15]

The recognition that institutions may be an important—and, possibly, the most vital—aspect of economic growth is the reason that there is now such elevated interest in the role of law in fostering development. Law is, after all, the institution that may be the one most under human control and the most amenable to change (by comparison to, say, cultural and religious institutions).

The interest in the role of institutions in furthering economic growth also had a great impact on the methods by which international organizations extended aid to developing countries. For example, the 'Washington consensus' of the early 1990s had as one of its premises that aid should be conditional on the recipient's adopting institutions that were thought to be conducive to modern economic growth—such institutions (and policies) as fiscal discipline, privatization of state-owned enterprises, trade liberalization, and deregulation. There is now general agreement that the reform experiences of the 1990s, conducted under the Washington consensus, were not successful.[16]

The scholarly literature on the causes of economic growth and the practical policy advice that might follow from it are still evolving. Even so, there are two fundamental points about the causes of growth

[15] For a sceptical view of the role of institutions, see Edward Glaeser, Rafael La Porta, & Florencio Lopez-de-Silanes, and Andrei Shleifer (2004), 'Do Institutions Cause Growth?,' *J. Econ. Growth*, 9(271).

[16] See World Bank (2005), *Economic Growth in the 1990s: Learning from a Decade of Reform*. See also Dani Rodrik (January, 2006), 'Goodbye Washington Consensus; Hello Washington Confusion?' (available at http://ksghome.harvard.edu/~drodrik/Lessons%20of%20the%201990s%20review%20_JEL_.pdf (last accessed on 17 December 2009). At the conference in Hyderabad at which this paper was first presented, Hans-Bernd Schäfer commented that the Washington consensus was not merely neutral (in the sense of doing no harm even if it did not do much good): it actually did harm in some areas of the world, notably Latin America and Russia (both of which implemented 'big bang' policies to move to more 'Washington-consensus friendly' policies). Bernd noted that two dire consequences of the implementation in Russia were that life expectancy went from 63 to 55 and per capita income fell dramatically. On the bad effects of the adoption of 'Washington-consensus friendly' policies in Latin America, see Francis Fukuyama (2008), *Falling Behind: Explaining the Development Gap Between Latin America and the United States*.

about which there is a general consensus. The first point is that there is no persuasive evidence that the more successful countries achieved their growth by exploiting other countries. That is, modern economic growth was *not* the result of a systematic transfer of resources from the poor to the rich. The rich came by their wealth relatively honestly (although we are not precisely sure how they managed it). There is great hope in this observation: economic growth is not a zero-sum game in which some are destined to prosper and others not; it is not the case that the only viable path to sustained growth and prosperity is to take from others, thereby impoverishing them until they have the strength to take from you. Rather, the hopeful pattern is, to paraphrase Deng Xiao-ping, that 'All will be wealthy; some will be first.'[17]

A second point well worth making it that among the many possible causes of growth, one that almost no serious student of the process takes seriously is that some portions of humanity—races, religions, ethnic groups—are more adept at growth than are others.[18] Economic growth, prosperity, and happiness are achievable by all humans. This hopeful point should not be taken to undermine the earlier point about the importance of institutions: all human societies may grow

[17] I am grateful to Hans-Bernd Schäfer for this paraphrase. I do not mean to be Pollyanna-ish about the prospects for growth. Some countries and regions have greater resource endowments than others; some have greater access to trade than others; and so on. Nonetheless, if it is institutions that are more important than geography, the prospects for meaningful growth are good everywhere.

[18] There is an important strain in the economic development literature that has held that 'culture' is an under-appreciated force in economic growth and development. Max Weber (1904), for example, in *The Protestant Ethic and the Spirit of Capitalism*, argued that northern Europe developed before southern Europe because the prevalent Protestant religions of the north urged its adherents to work hard in this life so as to reveal through their degree of worldly success the extent to which they were God's chosen. In contrast, Weber argued, the fatalistic theism of the Catholic religion prevalent in southern Europe created no incentive for hard work in this life. A more modern proponent of the view that 'culture matters' is the great economic historian David Landes. In his (1999) *Wealth and Poverty of Nations: Why Some Are So Rich and Some So Poor*, Landes argues that culture is the most important factor in explaining and predicting national prosperity. He notes, as have many others, that it is not mere coincidence that the countries that are the wealthiest have a connection to English culture. See also his (2001) 'Culture Makes Almost All the Difference,' in Lawrence E. Harrison and Samuel P. Huntington (eds), *Culture Matters*. Finally, see Francis Fukuyama (1995), *Trust: The Social Virtues and the Creation of Prosperity*.

so long as they have growth-favouring institutions. The reconciliation of the points may lie in the observation that all human societies are capable of finding and implementing those institutions.

11.2.3 More Recent Experience

As I indicated above, roughly all countries began the modern period of growth at the same level of well-being. But they did not end at the same level. Indeed, one of the most significant results of this 180-year period of modern economic growth is the fact that the gap between the richest and the poorest has widened throughout that period. Today the richest country in the world, the United States, has a per capita income that is approximately fifty times greater than that of the poorest country.

The clearest-eyed view of the current state of comparative development holds that *all* nations follow the same general path from underdevelopment to sustained growth but that (as Deng Xiao-ping saw) some began a long time ago and others have begun the process at later dates.[19] For example, the leaders (the United States, Canada, and the countries of northern Europe) began the process in the late eighteenth or early nineteenth century.[20] Mexico began the process in

[19] Stephen L. Parente and Edward C. Prescott (2005), 'A Unified Theory of the Evolution of International Income Levels', in Philippe Aghion and Stephen Durlauf, (eds), *The Handbook of Economic Growth*. See also Stephen L. Parente (2008), 'Narrowing the Economic Gap in the 21st Century,' in Kim R. Holmes, Edward J. Feulner, and Mary Anastasia O'Grady, (eds), *2008 Index of Economic Freedom*, available at http://www.heritage.org/Index/ (last accessed on 21 October 2009).

[20] Sachs (2006), pp. 33–40. Sachs asks, 'Why was Britain first?' And he gives six reasons for Britain's being a leader in modern economic growth. (1) 'British society was relatively open, with more scope for individual initiative and social mobility than most other societies of the world.' At 33. (2) 'Britain had strengthening institutions of political liberty. Britain's parliament and its tradition of free speech and open debate were powerful contributors to the uptake of new ideas. They were also increasingly powerful protectors of private property rights.' Id. (3) 'Britain became one of the leading centres of Europe's scientific revolution.' Sachs counts this factor as being critically important. Ibid. (4) 'Britain had several crucial geographical advantages. It had low-cost sea trade routes with most of Europe and proximity to North America, where the 'new settlements ... provided vast new territories for food production and raw materials such as cotton for British industry.' Those settlements also provided a 'safety valve that facilitated

the mid-nineteenth century. Japan began with the Meiji Restoration of the 1860s. Brazil began in the 1920s. The 'late starters', such as India, China, and some sub-Saharan countries, have initiated the growth process only since World War II. There are good reasons for believing that the experience of the leaders will be replicated among the late starters, subject, of course, to the fact that there are different resource and other constraints, different technologies, and different policy concerns (such as global warming) facing the late starters than those faced by the leaders when they began the process.[21] Indeed, it might be the case that getting a late start may have some advantages. For example, rather than work their way through the various technologies of communication in the same order that the leaders have done, late starters can go straight to the mobile phone and the dizzying array of communications possibilities that the near future holds out.

Do we know what the state of development is in the countries that are and have been seeking to grow? Yes, and the record is somewhere between mixed and encouraging. The 'mixed' part is that there are some identifiably poor regions of the world—that is, serious difficulties with growth appear to be confined now to only certain areas rather than to be a pervasive and nearly worldwide phenomenon. To the extent that we now know where the worst poverty is—in the central Asian republics, some of the twenty-two Arab-speaking countries, and sub-Saharan Africa—we can target aid and other policies designed to address the particular problems of those areas.

The 'encouraging' part of the current record is that there is evidence to suggest that the world distribution of income is improving—that is, that the relative and absolute numbers of the poorest people in the world has been declining and declining rapidly. Xavier Sala-i-Martin of Columbia University has argued that 'the proportion of the world's people living in acute poverty (on less than a dollar a day) fell from

the exodus of impoverished people from the British countryside.' Ibid. (5) 'Britain remained sovereign and faced lesser risk of invasion than its neighbors.' At 34. And (6) 'Britain had coal'. At 35. Sachs concludes his enumeration of these advantages with this marvellous observation: 'When a society is economically dominant, it is easy for its members to assume that such dominance reflects a deeper superiority—whether religious, racial, genetic, cultural, or institutional—rather than an accident of timing or geography.' At 39.

[21] Parente and Prescott, (2005).

17 per cent in 1970 to 7 per cent in 1998; the proportion living on less than 2 dollars a day fell from 41 per cent to 19 per cent.'[22] Sala-i-Martin further contends that the absolute number of people in the world who were living at the lowest level of wellbeing (one dollar per day or an annual income of 500 dollars) fell by 200 million over the same period and that the absolute number of people living at the intermediate level of poverty (two dollars per day or an annual income of 1,000 dollars) also fell by 350 million over the same period.[23]

This finding is controversial. There are obvious measurement problems having to do with the validity and comparability of statistics from different countries. But the biggest controversy has to do with the underlying issue of whether to compute poverty using national income statistics (as Sala-i-Martin does) or using household surveys of consumption (as others have done).[24] Calculation of poverty levels by Chen and Ravallion at the World Bank using household survey data shows little improvement in world poverty in the 1990s (and beyond).[25] They find that the percentage of world population living on less than a dollar a day in 1987 was 28 and that that figure had fallen by 1998 to only 24 per cent of world population. In contrast, Sala-i-Martin,

[22] See Sala-i-Martin (2006), 'The World Distribution of Income: Falling Poverty And … Convergence, Period,' *Q.J. Econ.*, 121(351). The quotation is from a summary of an earlier version of Sala-i-Martin's work (11 March 2004) in 'More or less equal,' *The Economist*.

[23] Some commentators, such as Surjit Bhalla (2002) in *Imagine There's No Country: Poverty, Inequality, and Growth in the Era of Globalization*, argue that these data indicate that the UN Millennium Development Goal of reducing the number of people living on dollar a day or less to half its 1990 level by the year 2015 had already been achieved by the time that goal was announced.

[24] There is no consensus among those learned in these matters about which source of data is preferable.

[25] See Shaohua Chen and Martin Ravallion (2001), 'How Did the World's Poor Fare in the 1990s?,' *Rev. Inc. & Wealth* 47(283), and Chen and Ravallion (2004), 'How Have the World's Poorest Fared Since the Early 1980s?,' *World Bank Res. Obs.*, 19(191). The authors have updated their work to the present and have found similar discouraging results in Chen and Ravallion (August, 2008), 'The Developing World Is Poorer Than We Thought, But No Less Successful in the Fight Against Poverty,' World Bank, Development Research Group, Policy Research Working Paper 4703, available at http://www-wds.worldbank.org/external/default/WDSContentServer/IW3P/IB/2008/08/26/000158349_20080826113239/Rendered/PDF/WPS4703.pdf (last accessed on 21 October 2009).

using national income data, had found a decline to 7 per cent of world population and a decline in the absolute number of people in poverty.

Another issue of some moment in this matter is the extent to which rapid growth in China and India is giving a false impression of worldwide trends. Those two countries account for 40 per cent of the world's population and, till recently, for a sizeable portion of the world's poorest people. The nearly 10 per cent per year growth rates over the past almost 20 years in those two countries have lifted very large numbers of people in China and India out of the direst poverty. Could it be that those successes blind us to a failure of improvement in other parts of the world?

Finally, some commentators note that while income inequality may be lessening across all poor nations, it may be widening among all nations—that is, while there may be improvement among the poorest world population, the rich are getting richer at a faster rate. And this trend might be replicated within the nations that are making progress in improving the well-being of the worst off. That is, even within the poorest countries, the income of the well-to-do could be increasing at a faster rate than that of the very poorest so that the degree of inequality within the poorest nations may be worsening. These critics ask then what sense it makes to be cheerful about the reduction in the proportion or absolute number of the poorest people in the world if both the gap in income between the richest and the poorest nations, and the gap in income between the richest and the poorest within nations have widened.

These are vexing issues to which there appear to be no simple answers. Perhaps the most appropriate stance is to be cautiously optimistic about improvement in the condition of the very poorest but not to relax further efforts to continue those improvements. One encouraging way to look at recent developments is this one from Jeff Sachs, 'Approximately 4.9 billion people live in countries where average income—measured by GDP per person—increased between 1980 and 2000. An even larger number, roughly 5.7 billion people, live in countries where life expectancy increased.'[26]

There is one sure method to improvement on which all thoughtful commentators agree: reduce population growth in the very poorest

[26] Sachs (2006), p. 51.

countries. There is no doubt at all that those countries that have managed to reduce their fertility rates are the ones that are either wealthy or are experiencing sustained economic growth. Half of the roughly 200 nations in the world, including the entire developed world, is at the zero population growth level of two children per couple, one of which children is female. By contrast, the poorest countries in the world have fertility rates of five or more, with the result that national populations are doubling every 20 or so years.[27] Of particular note in this regard are the poor countries that have made the demographic transition from high fertility rates to the lower rates that characterize the wealthy countries. And note how rapidly these changes have occurred. Bangladesh, for example, reduced its fertility rate from 6.6 in 1975 to 3.1 in 2000. Iran reduced its fertility rate from 6.7 in 1980 (just after the Islamic Revolution of 1979) to 2.6 in 2000.[28]

There is a chicken-and-egg problem with fertility declines: will higher growth cause a fertility decline or will an earlier fertility decline lead to higher economic growth? Whichever the direction of causation or whether the changes are complementary, a demographic transition is an almost essential part of national growth and development. There is evidence to suggest that a global demographic transition has begun. China, the most populous country, has already begun that transition. And there are worries in some developed countries that there will not be a sufficient labour force in the future. Japanese newspapers have reported that, at current trends, there will be no Japanese people in the year 2569.

11.2.4 The Limits of Modern Growth

I cannot leave this mixed and encouraging news about modern economic growth without mentioning the fact that there may be limits to growth and that those limits may be real, binding, and very, very troublesome.

[27] Ibid., p. 64. Because the growth rate of GDP per capita is equal to the growth rate of GDP minus the growth of population and because increasing GDP per capita is the most important goal of development, high growth rates of population work at cross purposes with policies designed to increase well-being.

[28] Ibid.

One of those limits may be imposed by global climate change. The problem is that development, under our current technologies, involves a nation's using increasing per capita amounts of fossil fuels. Burning coal to generate electricity and using refined petroleum to provide transportation are putting ever-increasing amounts of carbon dioxide (CO_2) (and other greenhouse gases) into the atmosphere. And that CO_2 traps heat from the Sun close to the Earth's surface, leading to a rise in the average temperatures on the planet. There is a clear positive correlation between the concentrations of CO_2 in the atmosphere and Earth's average temperature. Since the dawn of modern growth in the late eighteenth century till today, concentrations of CO_2 have risen from approximately 275 parts per million (ppm) to approximately 380 ppm. If current rates of growth in usage of fossil fuels continue and, indeed, increase through 2050, the concentrations of CO_2 are expected to reach 500 ppm. The results of those increases are somewhat in controversy, but all serious students agree that the rise in concentrations is anthropogenic and that it could have serious, unforeseen, and costly consequences for the entire planet.[29]

Adding to the problem of fossil-fuel use is that the Earth's ability to absorb CO_2 through forests and the oceans is either compromised or decreasing. Deforestation to provide additional living space is removing one of the most effective methods of 'carbon sequestration'.

Let me relate this issue more clearly to modern economic growth. The current population of the Earth is approximately 6.5 billion. Of those, about 1 billion live in developed countries where the annual per capita income is 20,000 dollars and above. About 1 billion live in poor countries where the annual per capita income is 1,000 dollars or less. The remaining 4.5 billion live in countries whose annual per capita income lies between 1,000 dollars and 2,000 dollars and that are experiencing sustained economic growth.

These broad generalizations hide the fact that the average consumer of fossil-fuel-related products and services in the wealthy countries annually consumes thirty-two times the quantity of those resources

[29] See, for example, Elizabeth Kolbert (2005), *Field Notes from a Catastrophe: Man, Nature, and Climate Change*. See also Nicholas Stern (2007), *The Economics of Climate Change: The Stern Review*. Eleven of the twelve warmest years in recorded history have occurred in the last twelve years.

than the average person in the developing world.[30] Put dramatically, each one of the over 300 million US citizens annually consumes as much fossil-fuel-related goods and services as 32 Kenyans. Presumably, those Kenyans and others in the 5.5 billion below US consumption standards aspire to emulate the consumption patterns or wealth positions of the average US citizen.

Imagine what would happen to the amount of fossil-fuel consumption if only one country, China, reached the US level of consumption. China has a population of approximately 1.5 billion, and per capita consumption rates of fossil-fuel-related goods and services there are only one-eleventh of what they are in the US (even though China has recently become, by volume, the greatest emitter of greenhouse gases in the world). Suppose, for the sake of argument, that China continues its remarkable growth of almost 10 per cent per year into the near future so that their annual per capita consumption levels rise to ours. That fact alone (and assuming no other changes in the world) would double world consumption rates of oil, electric power, metals, plastics, and other greenhouse-gas emitting products and services. If India, with a current population of one billion people, were also to grow so that their annual per capita consumption rates rose to current US levels, then world consumption rates would triple. Finally, suppose that the entire developing world of 5.5 billion people were to grow to US annual per capita consumption levels. World consumption rates would increase to eleven times their current levels. This would be the equivalent of the world's population increasing, at current consumption levels, to 72 billion people from its current level of 6.5 billion.[31]

The clear problem that global warming poses is whether increasing wealth and development is consistent with widespread concerns for the stability of the environment. We know that development improves lives considerably, but we now also know (or strongly suspect) that

[30] Much of what follows is based on Jared Diamond (2 January 2008), 'What Is Your Consumption Factor?', *The New York Times*.

[31] This frightening thought experiment assumes that there are no dramatic technological changes that replace 'carbon dioxide and other greenhouse-gas emitting' technologies. Of course, there may be such changes. Indeed, there almost have to be in order for widespread growth leading to current resource-consumption levels to succeed.

it may do so at a significant cost. A planet that in the year 2100 or beyond is, on average, 5°C warmer than it is today will cause climatic change that has large costs for large numbers of human beings. For example, sea levels may rise, causing extraordinary expenses of shoring up large coastal cities or large costs of relocating population and fixed capital away from the coasts; growing seasons are likely to change, making some areas of agricultural productivity inhospitable but opening up other areas (such as Siberia) to cultivation; droughts are likely to become more common in areas, such as the Sahel, that are already hard pressed; and the melting of snows and glaciers could threaten the water supplies of South and East Asia where 40 per cent of the human population lives.

The distressing aspect of these thoughts is that the problem is a worldwide one, and yet we do not have international institutions or agreements that address the issues. Nor is there any realistic prospect of creating such institutions or agreements. In the words of one careful student of global warming, 'This is the most serious problem to face humanity.'[32]

11.3 LAW AND ECONOMIC GROWTH

As we have seen, the most promising recent literature in economics on the causes of sustained modern growth focuses on institutions. In this section I draw a connection between 'institutions' and law. I shall ultimately conclude that however important institutions may be to growth, the aspects of law to which people frequently point as being vital to national development are unlikely to be as flexible or as crucial as some other factors that I seek to identify.

11.3.1 Overview

One of the shortcomings of the literature on institutions is its failure to be specific about just what counts as an 'institution.' A core understanding is that 'institution' refers to the practices and organizations of public life—the societal, regional, and local governance organizations; and the executive, legislative, and judicial organs of government. But the literature seems to spread out from that

[32] Daniel Farber (2008), 'Modeling Climate Change and Its Impacts: Law, Policy, and Science', *Texas Law Review*, 86(1655).

core to encompass the institutions of civil society,[33] private religious institutions, and commercial institutions. And even beyond those private institutions the literature sometimes cites the structure of the family and extended clan and social norms. We might further include 'culture,' that elusive but suggestive term that encompasses shared notions of right conduct or behaviour, of ethical and moral values, and of senses of beauty, euphonious music, and art.

So broadly conceived, the idea of 'institutions' seems to swallow up all categories of explanation and, in so doing, become an ultimately useless catch-all. We have to have a clearer and more exclusive notion of an 'institution' in order for that notion to be useful in explaining and predicting modern growth. As an important adjunct of this more precise notion, we ought to be able to explain how those institutions that favour growth come into being, why they finally appeared in northern Europe in the late eighteenth and early nineteenth centuries, why they have appeared only later in the 'late starters,' how inappropriate institutions give way to those that favour growth, and how, if at all, human intervention can consciously create growth-enhancing institutions.

This is a formidable list of questions and a demanding research agenda—both far beyond the scope of this essay. I want to look at a more limited subset of the possible meanings of 'institutions' in order to see the extent to which that meaning of the word is vital to the onset and continuation of modern growth.

The subset of 'institutions' upon which I want to concentrate are those gathered under the rubric 'law'. I want to be inclusive

[33] See, for example, the pathbreaking work on the vital role of civil society: Robert Putnam (1992), *Making Democracy Work: Civic Traditions in Modern Italy*. Putnam compared similarly sized towns in northern Italy, some of which were well-governed and others of which were not. His attempt to explain the differences in why democracy worked in some towns and not in others focused on the role of the institutions of civil society—the clubs, choir groups, soccer team fan clubs, commercial organizations, neighbourhood associations, and other informal groups that operated in the space between the families of the towns and their government and served to unify families and individuals. Where those intermediate civic organizations were present, democracy worked well; where they were absent or weak, democracy did not work very well. Putnam's hypothesis is that intermediate civic groups provide connections and cross relationships that are essential to a well-functioning democracy.

in my definition of the institutions of law. So, I mean that term to include both the organizations that make and enforce the law (the legislature, administrative agencies, executive, and judiciary; the private legal profession; and the police) and the substantive content of the law. I want to include the realms of private law that seek to govern relationships between private parties, such as contract, tort, and property law. I also want to include public law—criminal law, corporate law, bankruptcy law, and so on. In addition, I want to include the general governance structure of the society—whether there is meaningful democracy; whether political considerations are suitably under control or inappropriately intrude into decisions that would better be decided on principled, neutral grounds. Despite my attempt to focus on a limited set of institutions, this is a wide-ranging set.

Let me begin with a statement of the case that there is a connection between these legal institutions and modern growth. The broadest view holds that law is vital to national growth to the extent that it creates incentives for people to behave in a growth-enhancing manner. This means, among other things, harnessing individual self-interest to serve the public good, inducing people to 'do good by doing well'. Thus, a business organization must have the appropriate incentives to produce and market so as to make a profit but, in doing so, must respect the interests of their customers and employees in not being harmed unreasonably by the products they are purchasing or producing. And the production process must not result in unwanted and uncompensated pollution of the air or water. Those seeking to make agreements that will bind them for a period of time so as to secure a benefit from others similarly and reciprocally bound for a period of time must have confidence that those agreements will be kept or, if broken, result in adequate compensation. Those artists and inventors who seek to create new expressive works or new, useful, and nonobvious inventions (and in so doing, forego other profitable employment opportunities) should have reasonable expectations of being able to enjoy the economic benefits of their expressive and creative activities. Property owners should have reasonable security in their assets and know themselves to be free from unreasonable interference in the enjoyment and employment of those assets from other individuals, organizations, and the government.

There are, of course, many details to be added to this account, such as provisions for education of the populace, subsidies for basic scientific research, well-functioning financial intermediaries, an equitable taxation system, a reasonable health-care system, security for foreign investors, bankruptcy, and so on. But the core of the argument linking economic prosperity to legal institutions is the one in the previous paragraph.

One elaboration on this argument that is often left unstated but is implicit is that the governance structure of the system—including the judiciary who will decide private litigation—must be reasonably stable, inclusive, and productively and prudently responsive to the real events of the nation. Put in the negative, the government ought not to be cravenly corrupt.[34]

The key to implementing this view of the vital role of these legal institutions in fostering economic growth is the discovery of a mechanism for getting a low- or no-growth society to adopt these legal institutions. And so far, no obvious mechanism has presented itself. An important part of the Washington consensus was the attempt to condition aid from the international organizations on the adoption of some of the policies and institutions noted above. And as we have seen, that conditionality was not successful and, indeed, detrimental to sustained growth in several instances. As I also indicated above, we do not yet have a clear account of why these legal institutions—assuming for the moment that they *are*, in fact, essential to sustained modern growth—were adopted in northern Europe in the late eighteenth and early nineteenth centuries, nor why some former colonies (mostly former English colonies) saw fit to adopt those legal institutions while others did not. We do not know, in the parlance of comparative law, why some legal institutions are attractive as transplants and others are not. And because of our lack of knowledge about the reasons for the evolution or adoption of growth-enhancing legal institutions, we are at a loss as to how to persuade developing countries to follow the developed world's lead in this regard.

[34] There may be such a thing as 'optimal corruption'. The intuition is that some corruption is both inevitable and good and that attempts to zero-out corruption will be excessively costly and, among other social costs, may lead to discouraging otherwise talented people from devoting their energies to politics.

11.3.2 Why Good Law is Not Sufficient to Growth

The previous section laid out the argument linking law to the view that the correct institutions are vital for growth and also noted that we do not know how to implement this view in the slow developing countries or 'late starters.' In this section I want to express a deeper scepticism about the relationship between growth-enhancing legal institutions and modern growth.

My sense is that the arguments above about the vital role of private and public law and the legal profession are overblown and possibly inaccurate and misleading. This intuition takes some force from the evidence that the Washington consensus was not a successful growth plan. Nonetheless, my scepticism is not solidly grounded in either theory or thorough empirical evidence. As a result, these suggestions are preliminary and tentative, subject to revision in light of clearer theoretical and empirical arguments.

My arguments are three-fold. First, the view that law is central to growth ignores the fact that there are alternatives to law that can and do serve the same general functions as law. Second, the view ignores the great contribution that stability, in and of itself, can make to national growth and development. And third, the view underestimates the things that can go wrong in a society that seems to have a well-functioning legal system and also underestimates the value that a courageous political leader can have in fostering growth.

11.3.2.1 Alternatives to Law

The great psychologist Abraham Maslow memorably said, 'If the only tool you have is a hammer, every problem is a nail.' Lawyers have law as their principal tool, and, as a result, lawyers see every problem as amenable to legal solution. This 'legal centrism,' as Robert Ellickson called it,[35] suggests that other forms of ordering society are either ineffectual, unjust, or uncontrollable, or all three. But we now know that those alternatives to law are not only possible but that in many instances are the driving force in social organization.

Bob Ellickson's study of social norms in Shasta County, California,

[35] Robert Ellickson (1991), *Order Without Law: How Neighbors Settle Disputes.*

made this point eloquently.[36] Ellickson discovered that the law on the books regarding damage done by cattle to private property was different in two halves of a large county in northern California. In one half, the cattle-owner could be held liable for damage done by his unsupervised cattle to another person's property—say, a homeowner's garden; in the other half, the cattle owner could not be held liable for similar damage. Ellickson's hope was that if the costs of bargaining between cattle-owners and other property owners were low in both halves of the county, then the same practices for dealing with cattle damage might prevail, regardless of the law. If so, this would be a real-world indication that the Coase Theorem is correct.[37]

What Ellickson found was that the practices were identical in the two halves of the country, but that the reason was not that the Coase Theorem was correct. Rather, cattle-owners and private property owners in the two halves were seeking to conform to a clear and dominant social norm of being a 'good neighbour'. That norm required that if a cow strayed onto your property, you should take steps to protect that cow—by, for example, moving your car out of your garage and moving the cow in—call the cow's owner to tell him that you have the cow, and feed and protect the cow until the owner could come to pick it up. You should not remonstrate with the owner about the costs you were incurring; nor should you seek to get compensation from the cow-owner for any minor damage that you might have suffered.

Of course, the 'good neighbour' norm placed duties on the cow-owner, too. He should apologize for the inconvenience, offer to compensate the protector for the damage and the costs of housing and feeding his stray cow, and seek to come to retrieve the cow as quickly as possible. But so long as the norm was being observed by both sides, there was no talk of or need for lawyers. Indeed, Ellickson found that people in Shasta County only resorted to the law when people (typically recent immigrants from another region of the country) failed to obey the 'good neighbour' norm.

[36] Robert C. Ellickson (1986), 'Of Coase and Cattle: Dispute Resolution Among Neighbors in Shasta County,' *Stan. L. Rev.*, 38(623).

[37] The Coase Theorem holds that 'when transaction costs are zero, an efficient allocation will result, regardless of the law'. See, for an extended discussion, Robert D. Cooter and Thomas S. Ulen (2007), *Law and Economics*, 5th edn, pp. 85–99.

The central point of this famous example and of much subsequent work on similar topics is that for most people social norms, not law, are the principal governing rules of their lives. Law is invoked only in unusual situations where social norms have broken down or are not being followed.

How is this literature relevant to the role of law in economic development? The obvious connection is that social norms appear to be a close substitute for the law. Put a little more forcefully, law may not be necessary for a well-ordered society; social norms may do perfectly well. In fact, it may be the case that where the institutions of law are either absent or ill-managed, social norms become elaborate enough to perform many of the functions that a well-functioning legal system would have provided.

It may be, then, that the absence of a well-functioning legal system should not be taken as a signal that a society cannot sustain modern economic growth. It is possible that in developing societies, social norms are the principal guide to good behaviour, as they appeared to be in Shasta County, California, and that they perform well in guiding behaviour into efficient, growth-enhancing channels.

I am well aware that there can be detrimental social norms that impede growth as well as the beneficial social norms that substitute for law. Critics, such as Richard McAdams,[38] have noted, for example, that racism was a social norm in much of the United States that had a very toxic effect on society and required correction through concerted legal action. It is easily conceivable that developing countries or regions have deleterious social norms that prevent efficient investment or other growth-enhancing activities and that law might correct those adverse effects.

Nonetheless, a modern legal system may be a product of development, not a prerequisite to it. Until we know much more about the role of social norms in developing countries, we should not be so certain that social norms are not fulfilling the same societal functions that a well-functioning modern legal system might fill.

[38] Richard H. McAdams (1995), 'Cooperation and Conflict: The Economics of Group-Status Production and Race Discrimination', *Harv. L. Rev.*, 108(1003).

11.3.2.2 Stability and Growth

Decades ago Joseph Schumpeter characterized the capitalist system as one in which there was 'creative destruction'.[39] That wonderful phrase evokes the swirl of success and failure of new enterprises, the rise and fall of industries, the unsettling effects of new technologies that seems to many critics of capitalism to be chaotic and, therefore, undesirable. Modern growth brings along with it an unavoidable ferment and barely controlled chaos. There is no all-seeing institution in any society that can predict which enterprises should succeed and which should fail, which new technologies are going to have far-reaching effects, and which are going to stir minor ripples in the social and commercial fabric.

Nonetheless, stability is a valuable background condition for long-term growth. Investors, both foreign and domestic, must have reasonable assurances that next year will be much like this year. Too much uncertainty about fundamental conditions in which economic calculations are made will discourage anything other than very short-term decisionmaking. And that short-term decisionmaking is not conducive to the longer-range planning that seems such an essential part of modern growth.

How should a developing country seek to strike a sensible balance between allowing the ferment of modern growth and creating the stability within which that ferment may proceed? That, of course, is the key. And my primary intention here is to draw attention to this obvious and central point that may be lost in the literature on the salutary effects of good institutions and of good law—stability of background conditions is extremely important to long-run growth. Indeed, there may be a tension between having the correct legal institutions and having stable institutions such that stable-but-not-ideal institutions are more conducive to growth than less-stable-but-ideal institutions.

We do not know the relative merits of stability versus correctness. But we have a hint that there is this trade-off in the literature on the role of democracy in fostering economic growth. Democracy is, like sound legal institutions, thought in some quarters to be a necessary

[39] Schumpeter (1942), *Capitalism, Socialism, and Democracy*. See also Thomas K. McCraw (2007), *Prophet of Innovation: Joseph Schumpeter and Creative Destruction*.

precondition of modern growth. However, the great sociologist Seymour Martin Lipset noted in 1959 that democracy is typically a product of modern growth, not a precursor.[40] Because democracy can be an untidy system of government, societies just embarking on modern growth may prefer societal governance that is more stable than democracy and only tolerate the vagaries of democracy when the process of growth has taken such firm hold that stability of the background conditions (if not of the actual people in government) has been secured.

11.3.2.3 Political Courage

As noted, modern growth unsettles society. The swirl of 'creative destruction' raises some and dashes others; it elevates some sectors of the economy at the expense of others; it may expose previously secure domestic producers to withering competition from foreign enterprises; and so on. Accepting these and other changes is not at all easy. The prospect for betterment resulting from a change must be substantial for individuals to prefer the maelstrom of creative destruction to the less rewarding but comfortable life of low- or no-growth stasis. And most people prefer the known present to the unknown future, even with assurances of betterment. Change will command widespread approval or tolerance only if it will positively affect the vast majority of people and that betterment is far more likely than not.

Regardless of what form of government a developing nation has, navigating the politics of modernization is extremely taxing on political skills. Indeed, the social strains of modernization can result in rioting and rebellion.[41] But even if they do not, modernization

[40] The 'Lipset hypothesis,' sometimes called the 'Aristotle-Lipset hypothesis,' appeared in (1959) 'Some Social Requisites of Democracy, Economic Development, and Political Legitimacy', *Am. Pol. Sci. Rev.*, 53(69) and was elaborated in *Political Man: The Social Bases of Politics* (1960). For modern, empirical confirmation of the Lipset hypothesis, see Adam Przeworski, Michael E. Alvarez, Jose Antonio Cheibub, and Fernando Limongi (2000), *Democracy and Development: Political Institutions and Well-Being in the World, 1950–1990*. For a fascinating case study of this point, see Acemoglu, Johnson, and Robinson (2005), 'The Rise of Europe: Atlantic Trade, Institutional Change, and Economic Growth', *Am. Econ. Rev.*, 95(546).

[41] See, for example, Amy Chua (2002), *World on Fire: How Exporting Free-Market Democracy Breeds Ethnic Hatred and Global Instability*. See also Tom Ginsburg (2004), 'Democracy, Markets and Doomsaying: Amy Chua's *World on Fire*', *Berkeley J. Internat'l L.*, 22(310).

can place responsive politicians into very difficult situations. Growth involves change, and where there is change, there are gainers and losers. The gainers will surely encourage and accept the change, but the losers will fight to prevent their losses from becoming permanent or crippling. And because the losers are likely to have been people with considerable power in the pre-growth regime, they may begin their campaign to stall change from a very powerful position.[42]

There are two important implications of these observations. First, if change is disruptive and people are likely to resist it forcefully, is this not an argument for having a dictatorship or other form of nonresponsive government when modern growth begins?[43] This would certainly seem to follow, but I am sure that the suggestion in favour of dictatorship goes too far. It is true that a dictatorship could ignore the protests of those who are hurt by modernization or compensate them in some way. That is, the political costs of managing the transition to modernization might be lower under a dictatorship than under a responsive government. But the catch is that the dictatorship would have to be convinced that *it* would be better off under modern growth than under the alternative. And it is not at all clear that a dictator would always or even usually conclude that calculation in favour of modernization. Some, perhaps most, dictatorships will focus on what is best for them, not on what is best for the entire society. So, to be more precise, the argument in favour of overcoming political opposition to modernization's changes makes a case for an enlightened dictator, one who sees the benefits to all of the changes and subordinates his or her own desires in favour of that social good. Such dictatorships exist, but they are rare.

To be fair, it is not clear that a responsive government will opt for modernization either. It may mistakenly conclude that the costs of modern growth exceed the benefits.[44] But it seems more likely that a responsive government will respond in favour of sustained growth

[42] Daron Acemoglu and James Robinson (2000), 'Political Losers as a Barrier to Economic Development,' *Am. Econ. Rev.*, 90(126).

[43] I am grateful to Ajit Mishra for posing this important question.

[44] As I write this, the House of Representatives of the United States Congress has voted against the financial bailout plan proposed by Secretary of the Treasury Henry Paulson and supported by an overwhelming proportion of economists. So, even the world's oldest democracy can sometimes make questionable decisions.

when the benefits far exceed the costs. As I said above, there are good reasons for believing that a dictatorship will only take into account its own costs and benefits, not societal benefits.

Second, a further implication of the points above about the political costs of modernization is that nearly every society will opt for growth when the costs are small and the benefits large or when the gap between costs and benefits are very large and where the benefits will be widespread and when the likelihood of the benefits' being realized are high. But this is a formidable list of preconditions, and in real-world situations, these preconditions may either not exist or may be attended by a large degree of uncertainty. For instance, it may be the case that the benefits are distant and hazy, while the costs are vivid and immediate.

What is to be done in those circumstances? I believe that in those instances a society needs leaders with extraordinary courage, vision, and determination. They need vision to see the benefits of modern growth, courage to take on the costs of battling the political losers from modernization, and determination to continue the battle to victory. These are, needless to say, rare qualities in anyone and perhaps rarer still in professional politicians. Nonetheless, a developing society that can identify and encourage courageous politicians is likely to make the transition to modern growth quickly and with a minimum of pain.[45]

11.4 MODERN GROWTH IS SUSTAINED

This essay began with a review of the record of modern growth, noting that the experience of the past 180 years or so is fundamentally different from any preceding episodes of national wealth in human history in that this modern growth is sustained, not exploitative, and open to all nations and all humanity. I also briefly rehearsed the various academic theories of the causes of modern growth and how those theories have been instantiated in policies. I noted that there is an emerging consensus among students of growth and development

[45] In Hyderabad, Hans-Bernd Schäfer noted that politicians who are frequently credited with being great leaders—such as Nehru and Gandhi in modern Indian history—have in fact been disastrously wrong-headed from an economic point of view, while spectacularly successful leaders, such as Manmohan Singh, are frequently undervalued.

that 'institutions' are a vital ingredient, and I argued that among the various possible organizations and practices that the word 'institution' might encompass, 'law' and 'legal institutions' seem to occupy a core position in fostering national growth.

Nonetheless, I expressed some scepticism about the ability of legal institutions to be the 'golden key' to growth that has so long been sought. My scepticism centres on three factors—that there are suitable alternatives to growth-facilitating legal institutions and that many societies have found those alternatives, that just as important as the right institutions is social and national stability, and that focusing on the role of institutions and law as engines of growth draws attention away from the crucial role that political factors and particularly political courage can play in fostering a nation's growth and development.

REFERENCES

Acemoglu, D. (2008), *Introduction to Economic Growth*, Princeton University Press, Princeton, NJ.

Acemoglu, D. and J.A. Robinson (2005), 'Economic Origins of Dictatorship and Democracy', Cambridge University Press, Cambridge, UK.

—— (2000), 'Political Losers as Barriers to Economic Development', *American Economic Review*, vol. 90, p. 126.

Acemoglu, D., S. Johnson, and J.A. Robinson (2005), 'The Rise of Europe: Atlantic Trade, Institutional Change, and Growth', *American Economic Review*, vol. 95, p. 546.

—— (2002), 'Reversal of Fortune: Geography and Institutions in the Making of the Modern World Income Distribution', *Quarterly Journal of Economics*, vol. 117, p. 1231.

—— (2001), 'The Colonial Origins of Comparative Development: An Empirical Investigation', *American Economic Review*, vol. 91, p. 1369.

Bhalla, S., 2002, *Imagine There's No Country: Poverty, Inequality and Growth in the Era of Globalization*, Institute for the International Economics, Washington D.C.

Chen, S., and M. Ravallion (2008), 'The Developing World is Poorer Than We Thought, But No Less Successful in the Fight Against Poverty', World Bank, Development Research Group, Policy Working Paper 4703 dated August 2008; available at http://www-wds.worldbank.org/external/default/WDSContentServer/IW3P/IB/2008/08/26/000158349_20080826113239/Rendered/PDF/WPS4703.pdf (last accessed on 21 October 2009).

—— (2004), 'How Have the World's Poor Fared Since the Early 1980s?', *World Bank Research Observer*, vol. 19, p. 191.

—— (2001), 'How Did the World's Poor Fare in the 1990s?' *Review of Income and Wealth*, vol. 47, p. 283.

Chua, A. (2002), *World on Fire: How Exporting Free Market Democracy Breeds Ethnic Hatred and Global Instability*, Doubleday, NY.

Collier, P. (2007), *The Bottom Billion: Why the Poorest Countries Are Failing and What Can Be Done*, Oxford University Press, Oxford.

Cooter, R.D. and T.S. Ulen (2007), *Law and Economics*, Fifth Edition, Pearson Education Inc., N.J.

Diamond, J. (2008), 'What is Your Consumption Factor?', *The New York Times*, 2 January.

Easterly, W. (2002), *The Elusive Quest for Growth: Economists' Adventures and Misadventures in the Tropics*, MIT Press, Cambridge, Mass.

Ellickson, R. (1991), *Order Without Law: How Neighbors Settle Disputes*, Harvard University Press, Cambridge, Mass.

—— (1986), 'Of Coase and Cattle: Dispute Resolution Among Neighbors in Shasta County', *Stanford Law Review*, vol. 38, p. 623.

Farber, D. (2008), 'Modeling Climate Change and Its Impacts: Law, Policy, and Science', *Texas Law Review*, vol. 86, p. 1655.

Fukuyama, F. (2008), *Falling Behind: Explaining the Development Gap Between Latin America and the United States*, Oxford University Press, Oxford.

—— (1995), *Trust: The Social Virtues and the Creation of Prosperity*, Penguin, London.

Ginsburg, T. (2004), 'Democracy, Markets, and Doomsaying: Amy Chua's World on Fire', *Berkeley Journal of International Law*, vol. 22, p. 310.

Glaeser, E., R. La Porta, F. Lopez-de-Silanes, and A. Shleifer (2004), 'Do Institutions Cause Growth?', *Journal of Economic Growth*, p. 271.

Harrison, L.E. and S.P. Huntington (2001), *Culture Matters*, Basic Books, New York.

Helpman, E. (2004), *The Mystery of Economic Growth*, Harvard University Press, Cambridge, Mass.

Kolbert, E. (2005), *Field Notes From a Catastrophe: Man, Nature, and Climate Change*, Bloomsbury, U.S.A.

Landes, D. (1998), *Wealth and Poverty of Nations: Why Some Are So Rich and Some Are So Poor*, W.W. Norton and Company, New York.

Lipset, S.M. (1960), *Political Man: The Social Bases of Politics*, Johns Hopkins University Press, Baltimore, Maryland.

—— (1959), 'Some Social Requisites of Democracy, Economic Development and Political Legitimacy', *American Political Science Review*, vol. 53, p. 69.

Maddison, A. (2007), *The Contours of the World Economy 1-2030 AD: Essays in Macro-Economic History*, Oxford University Press, Oxford.

—— (2001), *The World Economy: A Millennial Perspective*, OECD, Paris.

McAdams, R.H. (1995), 'Cooperation and Conflict: The Economics of Group Status Production and Race Discrimination', *Harvard Law Review*, vol. 108, p. 1003.

McGraw, T.K. (2007), *Prophet of Innovation: Joseph Schumpeter and Creative Destruction*, Belknap Press of the Harvard University Press, Cambridge, Mass.

North, D.C. (2005), *Understanding the Process of Economic Change*, Princeton University Press, Princeton, NJ.

North, D.C. and R.P. Thomas (1976), *The Rise of the Western World: A New Economic History*, Cambridge University Press, Cambridge, UK.

Parante, S.L. (2008), 'Narrowing the Economic Gap in the 21st Century', in K.R. Holmes, E.J. Feulner, and M.A. O'Grady (eds) (2008), *Index of Economic Freedom, Heritage Foundation*; available at http://www.heritage.org/Index/ (last accessed on 21 October 2009).

Parante, S.L. and E.C. Prescott (2005), 'A Unified Theory of the Evolution of International Income Levels', in P. Aghion and S. Durlauf (eds), *The Handbook of Economic Growth*, North-Holland, Amsterdam.

—— (2002), Barriers to Riches, MIT Press, Cambridge, Mass.

Przeworski, A., M.E. Alvarez, J.A. Cheibub, and F. Limongi (2000), *Democracy and Development: Political Institutions and Well-being in the World, 1950–1990*, Cambridge University Press, Cambridge, U.K.

Putnam, R. (1992), *Making Democracy Work: Civic Traditions in Modern Italy*, Princeton University Press, Princeton, NJ.

Rodrik, D. (2006), *Goodbye Washington Consensus: Hello Washington Confusion?*, Kennedy School of Government, Harvard University; available at http://ksghome.harvard.edu/~rodrik/Lessons%20of%20the%201990s%20review%20_JEL__.pdf (last accessed on 21 October 2009).

—— (2004), 'Getting Institutions Right', April, Working Paper, Kennedy School of Government, Harvard University; available at http://ksghome.harvard.edu/~rodrik/ifo-institutions%20article%20_April%202004_.pdf (last accessed on 21 October 2009).

Rodrik, D., A. Subramanian, and F. Trebbi (2004), 'Institutions Rule: The Primacy of Institutions over Geography and Integration in Economic Development', *Journal of Economic Growth*, vol. 9, p. 131.

Romer, P. (2008), 'Economic Growth', in David R. Henderson (ed.), *The Concise Encyclopedia of Economics*, Liberty Fund, Indianapolis, USA.

—— (1990), 'Endogenous Technological Change', *Journal of Political Economy*, vol. 98, p. S71.

—— (1986), 'Increasing Returns and Long-run Growth', *Journal of Political Economy*, vol. 94, p. 1002.

Sala-i-Martin, X. (2006), 'The World Distribution of Income: Falling Poverty and…Convergence. Period', *Quarterly Journal of Economics*, vol. 121, p. 351.

Sachs, J. (2005), *The End of Poverty: Economic Possibilities of Our Time*, Penguin Press, New York.

Schumpeter, J. (1942), *Capitalism, Socialism, and Democracy*, George Allen and Unwin Publishers, U.K.

Solow, R.M. (1957), 'Technical Change and the Aggregate Production Function', *Review of Economics and Statistics*, vol. 312, p. 312.

—— (1956), 'A Contribution to the Theory of Economic Growth', *Quarterly Journal of Economics*, vol. 65, p. 65.

Spence, M. (2008), *The Growth Report: Strategies for Sustained Growth and Inclusive Development*, Chair, Commission on Growth Development; available at http://www.growthcommission.org/index.php?option=com_content&task=view&id=96&Itemid=169 (last accessed on 21 October 2009).

Stern, N. (2007), *The Economics of Climate Change: The Stern Review*, Cambridge University Press, Cambridge, U.K.

'More or Less Equal', *The Economist*, 11 March 2004.

Weber, M. (1904), *The Protestant Ethic and the Spirit of Capitalism*, English version published by Routledge Publishers, New York.

World Bank (2005), *Economic Growth in the 1990s: Learning from a Decade of Reform*, Washington D.C.

CONTRIBUTORS

P.G. Babu, Associate Professor at the Indira Gandhi Institute of Development Research (IGIDR), Mumbai and Adam Smith Guest Professor of Philosophy and Economics at University of Bayreuth, Germany.

Robert D. Cooter, Herman F. Selvin Professor of Law and Director of the Programme in Law and Economics at the University of California Berkeley School of Law.

Petra Ebermann, Research Assistant at the Institute of Law and Economics, University of Hamburg.

Thomas Eger, Professor of Law and Economics at the Faculty of law and at the Department of Economics, University of Hamburg.

Michael G. Faure, Professor of Comparative Private Law and Economics at the law faculty of the Erasmus University Rotterdam and Professor of Comparative and International Environmental Law at the law faculty of Maastricht University, the Netherlands.

Satish K. Jain, Assistant Professor at the Centre for Economic Studies and Planning, School of Social Sciences, Jawaharlal Nehru University, New Delhi.

A. Jayagovind, WTO Chair Professor at National Law School of India University, Bangalore.

Alexander S. Kritikos, Professor of Economics at the Hanseatic University Rostock, DIW Berlin.

Ajit Mishra, Senior Lecturer at the Department of Economics and International Development, University of Bath.

R.V. Ramana Murthy, Reader at the Department of Economics, University of Hyderabad.

Siddik Rabiyath, Research Scholar at the Center for Development Studies, Thiruvananthapuram.

A.V. Raja, Professor of Economics at the University of Hyderabad.

Padmanabha Ramanujam, Assistant Professor at the Jindal Global Law School, O.P. Jindal Global University.

Hans-Bernd Schäfer, Professor Law and Economics at the Bucerius Law School, Hamburg.

T.S. Somashekar, Assistant Professor at the National Law School of India University, Bangalore.

Thomas S. Ulen, Director of the Illinois Programme in Law and Economics and Professor of Law at the University of Illinois, College of Law, Urbana-Champaign, Swanlund Chair.